U.PORTO

FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

MSc

2.º CICLO

FCUP
2014

U.PORTO

Understanding the genetic basis of seasonal coat color change in the snowshoe hare, *Lepus americanus*: an RNA sequencing approach

Ana Mafalda Sousa Ferreira

FC

U.PORTO

FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Understanding the genetic basis of seasonal coat color change in the snowshoe hare, *Lepus americanus*: an RNA sequencing approach

Ana Mafalda Sousa Ferreira

Dissertação de Mestrado apresentada à
Faculdade de Ciências da Universidade do Porto em
Biodiversidade, Genética e Evolução

2014

MSc

Understanding the genetic basis of seasonal coat color change in the snowshoe hare, *Lepus americanus*: an RNA sequencing approach

Ana Mafalda Sousa Ferreira
Mestrado em Biodiversidade, Genética e Evolução
Departamento de Biologia
Centro de Investigação em Biodiversidade e Recursos Genéticos (CIBIO)
2014

Orientador
José Melo-Ferreira, Post-Doctoral Researcher, Centro de Investigação em Biodiversidade e Recursos Genéticos (CIBIO)

Co-orientadores
Paulo Célio Alves, Professor Associado com Agregação, Centro de Investigação em Biodiversidade e Recursos Genéticos (CIBIO) e
Faculdade de Ciências da Universidade do Porto
Jeffrey Good, Assistant Professor, Division of Biological Sciences, University of Montana

U.PORTO

FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Todas as correções determinadas pelo júri, e só essas, foram efetuadas.

O Presidente do Júri,

Porto, _____/_____/_____

“Snow has been such a constant feature of their environment that many northern animals have become well adapted to it and depend on it. Perhaps none depends on snow more than the snowshoe hare. (…) it would not take long for the predators to deplete them if it were not for their camouflage. The snowshoe (…) changes from a brown to a coat of pure white fur in winter. However, the more effective the camouflage is for one season, the less effective for another and a hare’s trick to survival requires getting the timing just right. (…) off-color hares are the first to be eaten and have their meat be reconverted into the next life, that of predators”

Bernd Heinrich, *in* “Winter World”

Agradecimentos

Os ciclos são próprios da Natureza. Tudo começa e tudo termina e como diria Mufasa: "Estamos todos ligados no grande Ciclo da Vida". Deparo-me agora com o culminar de um trabalho que marca o fecho de mais um ciclo na minha vida mas antes que ele se encerre não posso deixar de agradecer às pessoas que fizeram parte deste ciclo e que a ele estiveram ligadas.

Ao meu orientador, Doutor José Melo Ferreira agradeço toda a orientação e paciência, todas as respostas prontas a todas as dúvidas e mais algumas, a preocupação e a disponibilidade e por ser o melhor orientador que podia pedir. Ao meu coorientador, Professor Paulo Célio Alves agradeço por ter visto em mim o potencial para integrar o projeto onde se enquadra este trabalho o que me permitiu ver outro lado do Mundo que não conhecia. Obrigada por me entusiasmar com o seu entusiasmo. Obrigada aos dois por apostarem na minha formação e por acreditarem que eu seria capaz de fazer este trabalho.

To my co-supervisor Professor Jeffrey Good, I thank for making me feel welcome at the Good Lab, for the guidance and support and for always pushing me to do my best. To Professor Scott Mills and Marketa Zimova I thank for providing the samples and for their previous work on such an interesting system. Thank you to Colin Callahan for performing the extractions and library preparations. I also thank to Matthew Jones for sharing the same enthusiasm as I have and for the work he is developing on this system. To Thomas Brekke and Ted Cosart for teaching me some of the skills without which I would not be able to perform this work.

Ao Fernando Seixas e à Liliana Farelo por me terem acolhido no Grupo das Lebres.

I thank to the Good Lab in general for making me feel welcome from the first day. To Tom for Cosmic Encounters, to Ted for trading his knowledge for some Portuguese words, to Colin, Sara and Erica for always inviting me to Trivia, to Ryan my desk mate, to the rainbow boy Matt and finally to Jeff for receiving me. Also I have to thank to Jema and Maggie for being the best house mates ever and for not letting me get lost in Missoula. And finally to everyone I met at Missoula for having made me feel at home.

A todos os meus companheiros da biblioteca e de trabalho por terem tornado todo o tempo passado no CIBIO o melhor possível. Pelo apoio nas horas mais desesperadas, pelas brincadeiras e pelas cusquices. Um especial agradecimento à minha parceira de crime, a Mariana, por alinhar comigo nos gifs. Ao Luís por ser o melhor dude de sempre. À Ana Filipa por ser sempre tão otimista. À Sílvia pelas novelas, pelas chamadas, pelos skypes, pela constante e permanente amizade mesmo com 9h de diferença. À Carolina, a minha irmã e a minha sanidade. Ao Zé, a minha coragem.

Finalmente, à minha família por todo o apoio durante este ano de viagens. À Beatriz por me puxar para a brincadeira. Ao meu irmão porque nada disto era a mesma coisa sem ele. Ao meu pai por ter sempre um sorriso guardado mesmo nos momentos mais difíceis. À minha mãe, pela inspiração e por todo o amor.

Por último, a W.W.

Financial support

This work was supported by the Fundação para a Ciência e a Tecnologia (FCT) research project (PTDC/BIA-EVF/115069/2009; partial funds by FEDER through COMPETE) and a post-doc grant to JMF (SFRH/BPD/43264/2008; partial funds by POPH-QREN from E. Social Fund). It was also supported by funds from the Fundação Luso-Americana para o Desenvolvimento (FLAD).

Sumário

Evidências dos efeitos das alterações climáticas na biodiversidade em todo o planeta criam, por um lado, um problema de conservação mas, por outro, uma oportunidade para estudar a evolução de fenótipos adaptativos que surgem em resposta às novas condições ambientais. Uma adaptação importante de espécies árticas/boreais a habitats periodicamente cobertos de neve é a mudança sazonal da cor da pelagem. A diminuição dramática do período em que existe neve no solo causada pelo aquecimento global porá em causa a sobrevivência destas espécies devido a um aumento do período em que existe um desfasamento entre a cor do habitat e a do animal, potenciando um aumento da predação e consequente decréscimo da capacidade de sobrevivência. Perceber a base genética deste fenótipo é por isso fundamental para estudar o potencial adaptativo destas espécies e determinar se irão conseguir adaptar-se.

A lebre americana (*Lepus americanus*) muda sazonalmente a pelagem para castanho no verão (muda de primavera) e para branco no inverno (muda de outono), havendo algumas populações que mantêm a pelagem castanha durante todo o ano. Esta espécie é por isso um excelente modelo para realizar análises comparativas de expressão diferencial e de genómica populacional com o objetivo de compreender a base genética e o potencial adaptativo deste fenótipo. Neste trabalho, apresenta-se uma primeira abordagem para a compreensão da base genética da mudança sazonal da cor da pelagem na perspetiva da expressão genética e usando sequenciação de alto débito de RNA. Biópsias de pele que representam três fases de muda da cor do pelo, “branco”, “muda” e “castanho” foram recolhidas durante a muda de primavera em cinco espécimes de lebre americana capturados em Seeley Lake, Montana, Estados Unidos da América. O RNA foi extraído e sequenciado em plataformas Illumina HiSeq. As sequências obtidas foram então usadas para construir *de novo* um *assembly* do transcriptoma da pele, que foi anotado e usado para realizar uma análise de expressão diferencial entre os tipos de tecido que representam as diferentes fases da muda. Isto permitiu identificar genes envolvidos na maquinaria genética por trás da mudança sazonal da cor da pelagem.

Depois de filtrar possíveis erros no *assembly*, 14568 componentes do Trinity (potenciais genes) foram retidos, correspondendo a 12210 anotações ENSEMBL de genes. Da análise de expressão genética foi possível identificar 943 potenciais genes

diferencialmente expressos entre as três fases de muda definidas, e as principais diferenças ocorrem entre o início e o final da mudança de cor, que são paralelas às mudanças observadas durante o ciclo de crescimento de pelo. As funções que se descobriu estarem significativamente enriquecidas no conjunto de genes diferencialmente expressos incluem resposta imunitária, morfogénese da pele e termos relacionados com a organização da componente celular como a organização do citoesqueleto. Estas categorias estão provavelmente relacionadas com mudanças fisiológicas que ocorrem durante o ciclo de crescimento do pelo. Um análise mais aprofundada a alguns genes relacionados com a morfogénese do folículo capilar revelou a expressão diferencial de genes que têm um papel importante na proliferação de queratinócitos, formação da medula do pelo, estratificação da epiderme e produção de queratina. Adicionalmente, genes específicos relacionados com a melanogénese (ex. *ASIP, MC4R*), transporte de melanossomas, (ex. *MYOVIIA*) e ritmo circadiano (*NR1D1, RORα, RBSLO1/KCNMA1*) estão diferencialmente expressos entre as fases de muda.

Estes resultados providenciam uma primeira imagem sem precedentes das principais mudanças de expressão génica que ocorrem durante o ciclo de muda e identificam genes que possivelmente estão envolvidos na mudança de cor da pelagem que ocorre na lebre americana. Este trabalho fornece assim as bases para estudos futuros que pretendam pintar um quadro detalhado de como genes e interações genéticas determinam este fenótipo, e abre a porta para o estudo molecular do potencial adaptativo deste fenótipo num contexto de mudança climática.

Palavras-chave

Mudanças climáticas; *Lepus americanus;* lebre-americana; mudança sazonal da cor da pelagem; transcriptoma; sequenciação de RNA; expressão génica;

Summary

Evidences for the effects of climate change in biodiversity worldwide are both a conservation concern and an opportunity to study the evolution of adaptive traits in response to the new environmental conditions. A remarkable adaptation of arctic/boreal species in habitats periodically covered with snow is seasonal coat color change. The dramatic decrease of days with snow on the ground due to global warming will challenge the survival of these species due to increased mismatch with the background and potentially increased predation. Understanding the bases of this phenotype is thus fundamental to study the adaptive potential of these species and ultimately to understand whether they will be able to adapt.

The snowshoe hare (*Lepus americanus*) undergoes seasonal coat color change from a brown summer coat (spring molt) to a white winter one (autumn molt), despite the existence of individuals in some local populations that retain their summer coat color year-round. This species is thus an excellent model to perform comparative analyses of differential expression and population genomics with the aim of understanding the genetic bases and adaptive potential of this phenotype. In this work, a first insight on the genetic basis of seasonal coat color change is provided using a gene expression perspective and an RNA sequencing approach. Skin tissues representing three molting stages, “white”, “molting” and “brown” were sampled in five specimens of snowshoe hare undergoing the spring molt captured in the Seeley Lake region of Montana, United States of America. RNA was extracted and sequenced in high throughput using an Illumina HiSeq platform. The obtained reads were used to produce a *de novo* assembly of the skin transcriptome, which was annotated and used to perform a differential expression analysis among tissue types representing the different molt stages. Finally, a gene ontology term enrichment analysis was performed to understand the functional categories of the differentially expressed genes. This allowed identifying genes involved on the genetic machinery underlying seasonal coat color change.

After filtering out possible assembly errors, 14568 Trinity components (putative genes) were retained, corresponding to 12210 ENSEMBL gene annotations. After gene expression analysis, 943 putative genes were found to be differentially expressed between the three defined stages of molt, and major transcriptional changes occur between the

onset and end of the color change, which are parallel with the changes observed in the hair growth cycle. Gene functions significantly enriched in the set of differentially expressed genes include immune response, muscle tissue and structure development, metabolism, appendage development, skin morphogenesis and terms related with cellular component organization such as cytoskeleton organization. These categories are likely related with physiological changes occurring during the hair growth cycle. A deeper analysis into some genes related with hair follicle morphogenesis revealed the differential expression of genes that play a role in keratinocyte differentiation, hair shaft medulla formation, epidermis stratification and keratin production. Furthermore, specific genes related with melanogenesis (e.g. *ASIP*, *MC4R*), melanosome transport (e.g. *MYOVIIA*) and circadian rhythms (*NR1D1, RORα, RBSLO1*) were found to be differentially expressed among molting stages.

These results provide unprecedented insights on the major transcriptional changes occurring during the molt cycle and identify genes possibly involved in seasonal coat color change in snowshoe hares. This work sets the bases for future studies aiming at painting a detailed picture showing how genes and gene pathways determine this phenotype, and opens the door to the molecular study of the adaptive potential of this phenotype in a context of climate change.

Key-words

Climate change; *Lepus americanus;* Snowshoe-hare; seasonal coat color change; transcriptome; RNA sequencing; gene expression

Index

List of Figures

List of Tables

List of Abbreviations

ASIP – Agouti Signal Peptide

BCV – Biological Coefficient of Variation

BSWS – Brown vs White

CAGE – Cap Analysis of Gene Expression

cDNA – complement DNA

CPM – Counts per Million reads mapped

cRNA – complement RNA

DE – Differentially Expressed

DGE – Digital Gene Expression

DHI – Dihydroxyindole

DHICA – DHI Carboxylic Acid

dTTP - Deoxythymidine triphosphate

dUTP - Deoxyuridine triphosphate

EST - Expression Sequence Tags

ESU – Evolutionary significant units

FDR – False Discovery Rate

FGF10 – Fibroblast Growth Factor 10

FGF7 – Fibroblast Growth Factor 7

FP – False Positives

FPKM – Fragments per Kilobase of exon per Million reads mapped

GLM – Generalized Linear Model

GO – Gene Ontology

GWAS - Genome-Wide Association Study

IGFBP5 – Insulin Growth Factor Binding Protein 5

KRT25 – Keratin 25

KRT27 – Keratin 27

KRT71 – Keratin 71

L-DOPA - L-dihydroxyphenylalanine

logFC – log Fold Change

LRT – Likelihood Ratio Test

MC1R - Melanocortin-1 receptor

MC4R – Melanocortin-4 receptor

mRNA – Messenger RNA

MSWS – Molting vs White

MT2 - Melatonin receptor membrane

MTIF – Microphthalmia associated transcripton factor

MYOVA – Myosin Va

MYOVIIA - Myosin VIIa

NB – Negative Binomial

NGS – Next generation sequencing

NR1D1 –Nuclear receptor Rev-erbα

PAM – Partitioning Around Medoids

PD – Pars Distalis

PLOD3 - Procollagen-lysine, 2-oxoglutarate 5-dioxygenase 3

qPCR – Quantitative Polymerase Chain Reaction

QTL – Quantitative Trait Loci

RBSLO1 also known as KCNMA1 - Calcium-activated potassium channel subunit α-1

RIN – RNA Integrity Number

RNA – Ribonucleic acid

RNA-seq – RNA sequencing

RORα – Retinoid-related orphan receptor alpha

RUNX1 - Runt-related transcription factor 1

RUNX3 - Runt-related transcription factor 3

SAGE – Serial analysis of gene expression

SCN – Suprachiasmatic nuclei

TCV – Technical Coefficient of Variation

TMM – Trimmed Mean of M-Values

TP – True Positives

TP63 – Tumor Protein p63

TYR – tirosinase

TYRP1 –Tyrosinase-related-protein-1

TYRP2 – Tyrosinase-related-protein-2

UDG - Uracil-DNA Glycosylase

αMSH - α-Melanocyte-stimulating-hormone

1. Introduction

The global concern about anthropogenic-induced climate change and its effects on biodiversity worldwide has brought together efforts of scientists, governments and other people to produce evidence both of the phenomenon and its effects on Earth's organisms. The anthropogenic production of greenhouse effect gases, such as CO_2, will perpetuate a change in world climate leading, for example, to increased temperatures, changes in patterns of rainfall, sea level rise (Solomon *et al.*, 2009) and acidification (Hoegh-Guldberg *et al.*, 2007), and reduction of snow cover (Pederson *et al.*, 2011). In this context, biodiversity worldwide faces new conditions that challenge their adaptive capacity and ultimately their survival.

Although some prospects are not optimistic and extinction is the fate for many species (Thomas *et al.*, 2004), with a consequent decrease in biodiversity and ecosystem disruption, the new selective pressures may promote new adaptations (Franks and Hoffmann, 2012). Several studies show that human disturbances are in fact causing more phenotypic changes than what would be expected by natural environment action alone (Hendry *et al.*, 2008). Also, shifts in species ranges and life cycle events (phenology) have been found to be globally consistent with patterns of climate change (Parmesan and Yohe, 2003; Parmesan, 2006). Changes in the patterns of seasonality are for instance inducing changes in the migratory behavior of birds (Jenni and Kéry, 2003; Bearhop *et al.*, 2005), sometimes with maladaptive consequences (Inouye *et al.*, 2000). It is also inducing adjustments to an enlarged breeding season by promoting an advancement in reproduction time (Réale *et al.*, 2003; Franks *et al.*, 2007) and shortening diapause (Bradshaw and Holzapfel, 2001) and hibernation (Inouye *et al.*, 2000; Adamík and Král, 2008).

1.1. The study of the genetic bases of adaptive traits

Climate change can thus be regarded as an evolutionary force that may induce species to change and adapt to new conditions by the emergence of new phenotypes. In this case, phenotypes better adapted to the new environmental context will thus be favored and tend

to be predominantly passed to the following generations (Darwin and Wallace, 1858). Although selection acts on phenotypes, these are determined by genotypes and therefore selection leads to change in the frequency of the underlying genotypes. Understanding how to link phenotypes to genotypes is thus a relevant question that needs to be addressed in the context of human induced habitat change.

1.1.1. The genetic basis of phenotypes

Genetic variation at different levels may underlie different phenotypes. The simplest model possible is to consider one single gene where mutations may affect protein-coding genes that confer a change in phenotype (see e.g. Nachman *et al*., 2003), and/or changes in non-coding regions that affect expression level of a gene rather than its coding sequence (see e.g. Huang and Kang, 2007). At this level, structural variation such as gene duplications can also be considered (Maroni *et al.*, 1987). Nevertheless, the strength of this type of genetic variation has to be contextualized within gene networks and biochemical pathways that ultimately lead to the expression of a trait (Dalziel *et al.*, 2009). Some phenotypes can result from the interaction of several proteins and therefore a single-locus-based genetic variation will only have an effect on the trait if the allele promoting the change has a very strong effect (Hoffmann and Willi, 2008). For example, the success of variation arising at pleiotropic genes involved in more than one pathway is restricted by the effect it has in each of the pathways (Papakostas *et al.*, 2014). Also, although the neutral theory of evolution states that new mutations are the main source of genetic variation (Kimura, 1984), standing genetic variation – variation that exists previous to the onset of a selective force – being immediately available has been hypothesized to be of great importance to provide a quick response to selective pressures (Barrett and Schluter, 2008). Finally, hybridization between closely related species can also be a source of new genetic variation (Anderson *et al.*, 2009).

Different techniques have thus been developed to reveal genotype-phenotype associations. For instance, if the phenotype under search had been studied before it is possible that a candidate list of genes can be advanced from the scientific literature. For example, color phenotypes have been studied using the well-known pigmentation pathway (Bennett and Lamoreux, 2003; Nachman *et al.*, 2003; Steiner *et al.*, 2007). When this

knowledge is not available other solutions may be adopted. Linkage based methods such as Quantitative Trait Loci (QTL) and Genome-Wide Association Studies (GWAS) assume that loci under selection will be physically linked to known genetic markers (Orr, 2005). However, these methodologies have some challenges such as being dependent on the linkage disequilibrium between markers and loci, on the resolution that the markers provide (Orr, 2005; Hoffmann and Willi, 2008) and often require the use of inbreed lines (see for example Atwell *et al*., 2010). Moreover, outlier based methodologies base their principle on the fact that strong directional selection will increase the frequency of loci associated with adaptive traits in a given population (Le Corre and Kremer, 2012). The methodology is based on scanning for genetic markers in groups of individuals with different phenotypes and searching for outliers that are highly differentiated between them and correlate with the phenotype variation, even though these signals may be confounded by demographic effects and population structure (Excoffier *et al.*, 2009). Finally, expression studies can also be used to assess regulatory changes in gene expression that may explain phenotype-genotype links (Stoughton, 2005).

1.1.2. The study of gene expression

Regulatory features and gene expression can also be linked to phenotypic changes that can be captured by looking at gene expression. The expression shifts can merely translate plastic responses (Hoffmann and Willi, 2008) but mutations on regulatory regions themselves can change the expression of genes (Huang and Kang, 2007). Furthermore, the phenotype of a gene is affected by the protein and RNA machinery that acts above the gene level to produce functional proteins (King and Wilson, 1975; Eddy, 2001; Huang and Kang, 2007). Gene expression studies can also be valuable as an exploratory approach to draw a list of candidate genes for a given trait (see for example Henning *et al.*, 2013; Marra *et al*., 2014; Zhou *et al*., 2014) and are particularly suitable to understand the genetic machinery behind cyclic and seasonally variable traits (Lin *et al.*, 2004).

Expression can be studied using hybridization or sequencing-based techniques. The first use microarrays with DNA probes onto what fluorescence cRNA or cDNA tags hybridize (Wang *et al.*, 2009). However, microarrays require a previous knowledge about what markers to screen for to construct the array and face problems with cross-hybridization

(Eklund *et al.*, 2006) and saturation of the fluorescent signal (Wang *et al.*, 2009). On the other hand, sequence based techniques also known as Digital Gene Expression (DGE), use sequences or tags representing sequences to measure gene expression. Examples are expression sequence tags (EST) (Bouck and Vision, 2007), serial analysis of gene expression (SAGE) (Velculescu *et al.*, 1995) and cap analysis of gene expression (CAGE) (Shiraki *et al.*, 2003), that differ mainly in tag size or region of the mRNA from where the tag is produced. The shortcoming of these three methodologies is that they rely on expensive Sanger sequencing and vector cloning and although it is possible to produce these types of markers for non-model organisms they are expensive (Wang *et al.*, 2009). Also, to determine the tag annotation is necessary to map them onto a reference genome what might be relatively difficult since the tags are a shorter representation of the original cDNA sequence (Bouck and Vision, 2007; Wang *et al.*, 2009). Nevertheless, ESTs have indeed been applied to study adaptive traits and can be used to draw a panel of genetic markers of genome scans, for instance (Bouck and Vision, 2007).

RNA sequencing (RNA-seq) is the next generation alternative for gene expression studies. The molecular reactions implemented by the different sequencing platforms exclude the traditional cloning step and allow simultaneous sequencing of DNA from different samples (Shendure and Ji, 2008; Metzker, 2010; Mardis, 2013). Consequently, this allows the production of unprecedented amounts of genomic level data at lower price and time, opening the door for non-model organisms and to complex traits (Stapley *et al.*, 2010). It is based on the sequencing of a population of RNA (or mRNA) by reverse transcription to cDNA, ligation of each fragment to an adaptor that can be specific for every sample and sequencing in a NGS platform generating a single-end (if the fragments are sequenced from one end) or paired-end (if sequenced from both ends) dataset of reads (Mortazavi *et al.*, 2008). Reads can then be aligned to a reference or *de novo* assembled transcriptome and the number of reads mapped to specific genetic regions will be the proxy for gene expression level (Mortazavi *et al.*, 2008). Because of the high throughput and coverage of the output, RNA-seq data has more power to detect lowly or highly expressed features and does not require a prior knowledge about the transcriptome, as opposed to microarrays (Wang *et al.*, 2009). Also it does not require sequencing with cloning vectors and reads are longer that EST, CAGE and SAGE tags, allowing the performance of more reliable assemblies and mapping (Wang *et al.*, 2009). RNA-seq generated data can be applied for *de novo* transcriptome assembly of transcriptomes (Mortazavi *et al.*, 2008; Cahais *et al.*, 2012; Henning *et al.*, 2013; Gui *et al.*, 2013; Hershkovitz *et al.*, 2013; Gayral *et al.*, 2013;

Vellichirammal *et al.*, 2014), genetic marker discovery (Zhou *et al.*, 2014), gene expression analysis to pinpoint candidate genes for important adaptive traits such as osmoregulation (Marra *et al.*, 2014), coloration (Henning *et al.*, 2013), immune response and adaptation to aquatic life (Gui *et al.*, 2013) or extreme adaptation in snakes (Castoe *et al.*, 2013). It also can be applied to study patterns of splicing (Merkin *et al.*, 2012), improve the knowledge about transcriptome structure (Nagalakshmi *et al.*, 2008), gene expression analysis of developmental timelines (Wang *et al.*, 2010) or spermatogenesis (Margolin *et al.*, 2014). The major drawback of RNA-seq is related with the large amount of data that require large storage space, computational power and specific bioinformatics pipelines (Wang *et al.*, 2009). Nevertheless, methodologies to obtain the best results with this type of data have emerged. Examples are RNA preservation, storage and extraction protocols for RNA-seq (Gayral *et al.*, 2011), different quality trimmers (Del Fabbro *et al.*, 2013), *de novo* assembly optimization strategies (Martin and Wang, 2011; Duan *et al.*, 2012; Cahais *et al.*, 2012; Singhal, 2013), the different available assembly algorithms (Miller *et al.*, 2010), the performance of different methodologies for isoform detection and gene expression evaluation (Steijger *et al.*, 2013; Angelini *et al.*, 2014) or the different programs or pipelines used during the overall process (Vijay *et al.*, 2013; Wolf, 2013).

1.2. The adaptive role of color phenotypes

An important and well known adaptive trait is coloration and color patterns of hairs, scales, skin or feathers, and several adaptive hypothesis have been delineated to explain the role of coloration in Nature (Protas and Patel, 2008). One of them is communication for example in sexual interactions (Rowland *et al.*, 1991; Cuadrado, 1998; Ibáñez *et al.*, 2014). Contrasting tail or ear tips seem also to be important for communication for example in mammal carnivores (Ortolani, 1999) and lagomorphs (Stoner, 2003). Moreover, color is thought to be also important for thermoregulation with dark bodies allowing higher body temperatures or dark patches having an antiglare function, for example (Forsman, 1997; Ortolani, 1999; Hegna, 2013). Finally, coloration plays an important role in pray-predator interactions since it can either serve as a warning to predators or concealment. On the one hand, some species advertise their toxicity by displaying colors that make them conspicuous to predators such as *Heliconius* butterflies (Parchem et al., 2007 and references therein). Color can also be used to mimic other

species advertisements to trick predators into thinking they are toxic as well (Kraemer and Adams, 2014). On the other hand, the fitness cost of the disruption of crypsis between animals and their background is high because it will unveil their location either to prey or predators, and thus color phenotypes that allow camouflage tend to be favored by natural selection. Matching coat color with the background has been described in the rock pocket mice (*Chaetodipus intermedius*) were two color morphs exist, one light-colored that lives in sandy habitats and a melanic form that lives in lava fields (Nachman *et al.*, 2003). Similarly, two subspecies of oldfield mice (*Peromyscus polionotus*), *P. p. subgriseus* and *P. p. leucocephalusseems*, seem to have adapted to different colored habitats by evolving a dark coat in the mainland habitats and lighter coat in sand beach habitats (Steiner *et al.*, 2007). Spots and vertical and horizontal stripes are associated with disrupting the body shape in different habitats in mammals carnivores, with the first being associated with life in closed forests, vertical stripes with grasslands and horizontal stripes with canopies (Ortolani, 1999). Additionally, some animals change their color or color patterns according to the changing background such as the cuttlefish (Chiao and Hanlon, 2001) or according with the type of predator as in the case of the dwarf chameleon (Stuart-Fox *et al.*, 2006).

1.2.1. Seasonal coat color change

Another remarkable adaptation involving coat color is the seasonal coat color change in some mammals and bird that live in artic or temperature regions around the globe. Seasonal shedding of epidermal appendages such as feathers and hairs is common among vertebrates (Chuong, 1998), but these species are interesting because the color of their coating changes with season. Some examples are the rock ptarmigan (*Lagopus mutus*) (Jacobsen *et al.*, 1983), the Artic fox (*Alopex lagopus*) (Våge *et al.*, 2005), the Siberian hamster (*Phodopus sungorus*) (Logan and Weatherhead, 1978), the stoat (*Mustela erminia*) (Rust, 1965), the long-tailed weasel (*M. frenata*) (Bissonnette and Bailey, 1944), the collared lemming (*Dicrostonyx groenlandicus*) (Gower *et al.*, 1994), the white-tailed jackrabbit (*Lepus townsendii*) (Flux and Angermann, 1990), the mountain hare (*L. timidus*) (Angerbjörn and Flux, 1995), the Artic hare (*L. articus*) (Flux and Angermann, 1990), the Japanese hare (*L. brachyurus*) (Nunome *et al.*, 2014), the Alaskan hare (*L. othus*) (Best and Henry, 1994) and finally the snowshoe hare (*L. americanus*) (Severaid, 1945). Seasonal color change has a clear adaptive significance since it allows

concealment depending on the predominant color of the background, *i.e*, white in the winter that matches snow cover, or brown in the summer when snow melts. Given that it has been shown that the period of snow on the ground has been decreasing as a consequence of climate change (Pederson *et al.*, 2011), global warming may thus induce strong mismatches between coat color and the background, and endanger the survival of these species (Mills *et al.*, 2013).

1.2.2. The genetic basis of seasonal coat color change

The biochemistry, genetics and hormonal control of pigmentation in mammals has been studied and reviewed (Slominski *et al.*, 2004). Pigmentation in mammal's hair results from the accumulation of eumelanin (black to brown color) or pheomelanin (yellow to red color) inside melanosomes in the melanocytes of the hair bulb, which are then transferred to keratinocytes of the hair shaft (Slominski *et al.*, 2004). Eumelanin and pheomelanin are the end product of a pathway that starts with L-tyrosine (Slominski *et al.*, 2004) (see Figure 1.1). However, the color of the hair will depend not only on the type of melanin in the melanosome but also on the size and shape, number and distribution of the melanosomes deposited inside the hair shaft (Hearing, 1999).

Candidate genes for color morphs in the wild can be drawn from extensive literature on pigmentation patterns and colors described and studied in mammals, specially using mice as models (reviews in Bennett and Lamoreux, 2003; Hoekstra, 2006; Nakamura *et al*., 2002). Two of the most well-known candidate genes are the melanocortin-1-receptor gene, (*MC1R*) and the agouti signal peptide (*ASIP*) gene. For example, in the already mentioned example of the rock pocket mice, the color morphs have been shown to relate with mutations in melanocortin-1-receptor gene, *MC1R* (Nachman *et al.*, 2003). This MC1R receptor plays an important role in the change between eumelanin and pheomelanin production since its activation by α-melanocyte-stimulating-hormone (MSH) induces the production of eumelanin and its antagonist, the agouti signal peptide (ASIP), inhibits its function and leads to the production of pheomelanin (Voisey and Daal, 2002). A study in the Artic fox (*A. lagopus*) addressed the genetic basis of two winter color coats – gray and white – and two amino acid substitutions in *MC1R* have been implicated (Våge *et al.*, 2005). However, a recent study has failed to encounter a role for this gene's

polymorphisms in the genetic mechanism that explains the existence of populations that either do or do not change color in the winter in the willow grouse (*L. lagopus*) (Skoglund and Höglund, 2010).

Little is thus known about the genetic basis of the seasonal coat color change. However, the change likely involves a regulatory mechanism that prevents the accumulation of melanin in the hair without disruption of the mechanism that allows the production of melanin and consequently brown hair. Therefore, permanent mutations in genes that determine a permanent color phenotype may not be involved. For example, Logan and Weatherhead (1978) studied the seasonal variation in the concentration of tyrosinase (*TYR*) – the enzyme that takes part in the first steps of melanogenesis (Marmol and Beermann, 1996) – and melanin in the hair follicle of Siberian hamsters (*P. sungorus*) and observed that two peaks of tyrosinase are observed in spring and autumn, corresponding to the two molts observed in these hamsters. On the contrary, only one peak of melanin is observed and this occurs in spring, when brown hair is produced. No peak of melanin is observed in autumn, accordingly with the production of unpigmented hair which leads to the conclusion that, in this case, the pathway was disrupted between tyrosine and melanin production (Figure 1.1).

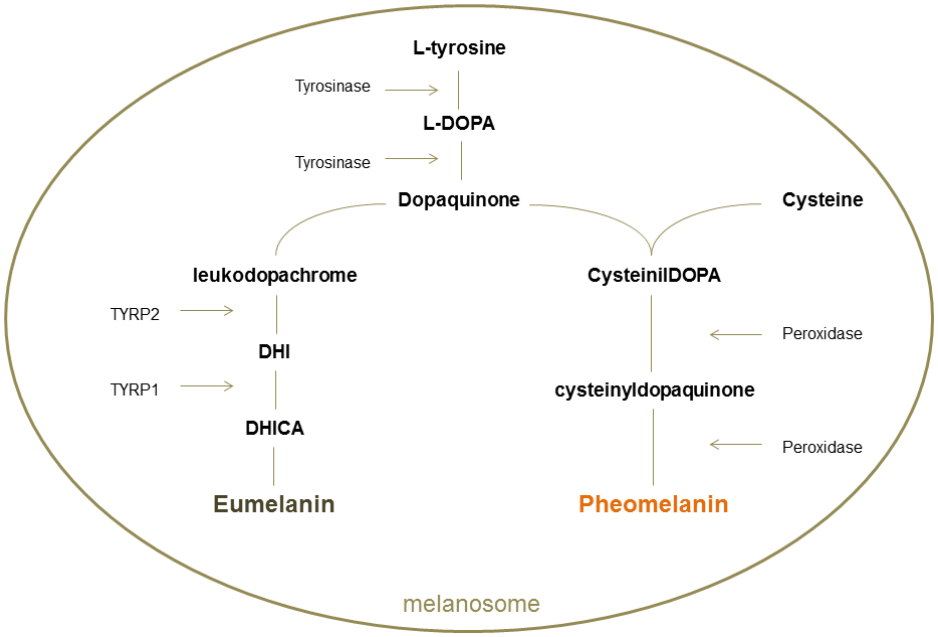

Figure 1.1 – A simplified scheme of melanogenesis pathway with the most important intervenients in the cascade. Abbreviates stand for L-dihydroxyphenylalanine (L-DOPA), Dihydroxyindole (DHI), DHI carboxylic acid (DHICA), Tyrosinase-related protein 1 (TYRP1) and Tyrosinase-related protein 2 (TYRP2). Based on Slominski *et al.* (2004).

As other seasonal phenotypes, such as reproduction, hibernation and migration, molts are regulated by a circannual rhythm controlled by the photoperiod (Goldman, 2001). In mammals, the integration of the photoperiodic signal is made by the pineal gland that receives that signal through the retina and suprachiasmatic nuclei (SCN) of the hypothalamus and produces a hormone, melatonin, in a rate inversely proportional to day length (Goldman, 1999). Melatonin is thus produced during the night and the longer the night period (shorter the day) the longer is the production of melatonin. Therefore, this production will change throughout the year and integrate the seasonal variation of photoperiod (Goldman, 2001). Melatonin then exerts a negative control over the Pars Distalis (PD) of the pituitary gland to mediate the production of prolactin; high levels of melatonin suppress the production of prolactin (Lincoln *et al.*, 2006). In Soay sheep studied by Lincoln *et al* (2006), the increased prolactin levels induced a molt in the pelage. Establishing the link between melatonin or prolactin production and hair follicle morphogenesis and pigmentation or seasonal coat color change has been attempted before. Sites for the ligation of melatonin on the murine skin have been found, especially in keratinocytes and epithelial bulb of the hair follicle (Slominski *et al.*, 1994). Kobayashi *et al*. (2005) described that the expression of melatonin receptor membrane (MT2) and retinoid-

related orphan receptor alpha (RORα) in murin and human cells coincide with the hair growth cycle, which suggest that they can be involved in the control of the cycle. Melatonin administration has been shown to inhibit tyrosinase (TYR) activity, and therefore melanogenesis, in a dose-dependent response and because only high doses have this effect, Slominski and colleges (1994) hypothesized that there is a non-identified ligand that mediates the process. The same dose-dependent relationship between melatonin and tyrosinase activity was observed in Siberian hamster (*P. sungorus*) (Logan and Weatherhead, 1980). An administration of α-MSH did not reverse this effect and despite the high levels of tyrosinase melanogenesis was still inhibited. This suggests a mechanism of inhibition that acts post tyrosinase in the pathway in Figure 1.1. Furthermore, prolactin injections in Siberian hamster inhibits the winter molt and are important to promote the onset of the summer molt (Duncan and Goldman, 1984). The action of melatonin was not inhibited in these hamsters and the injection of prolactin in hamsters kept at short-day photoperiod prevented the winter molt. Therefore, the action of prolactin seems to overlay the inhibitor action of melatonin on hair pigmentation described before. Also, a relationship with temperature has been established in Siberian hamster, with lower temperatures accelerating the appearance of the winter coat in hamsters kept in short days by increasing the responsiveness of melatonin target tissues (Larkin *et al.*, 2001). Despite this effect, variations in temperature do not explain completely the changes in the rate of molt and therefore it is not likely that it is the environmental cue triggering seasonal coat change (Mills *et al.*, 2013).

Although the knowledge about the hormonal regulation of seasonal coat color change is well established, the genetic basis of the mechanism is not known. Where does the melatonin/prolactin signal acts in the melanin pathway production? Although it seems that it acts beyond the step of tyrosinase production, there are several steps where the pathway can be interrupted. The inhibition could be on the production of melanin itself or on the transport of melanosomes to keratinocytes. Knowing the genetic basis behind this phenotype is vital to test its adaptive capacity to a changing environment and a scenario of mismatch and decreased fitness.

1.3. Seasonal coat color change in the genus *Lepus*

Genus *Lepus* spp., which includes hares and jackrabbits, is an excellent model to study the adaptive relevance of seasonal coat color. The genus has more than 30 described species that are widespread around the world and that occur in contrasting types of habitats from desert to tundra (Alves and Hackländer, 2008). Associations of coat colors to specific habitats, and consequently to specific background colors, has been studied in hares (Stoner, 2003). Generally, light color hares are associated with open habitats such as desert, tundra or barren land whereas dark colors are more frequent in forests or rocky habitats. Hares tend to display coats with shades of brown but some species living in high latitudes change their color to white in the winter which matches the snow covered ground. The seasonal coat color change phenotype has been described in species from Eurasia, such as *L. timidus* and *L. brachyurus*, and from North America, such as *L. othus, L. arcticus, L. americanus* and *L. townsendii.* Taking into account the phylogenetic analysis and biogeographic history of these species described by Melo-Ferreira *et al*. (2012), seasonal coat color change seems to occur in more than one evolutionary clade, and may thus have arisen independently. Furthermore, it is interesting to note that at least *L. timidus, L. americanus* and *L. brachyurus* have populations of individuals that retain summer coat color in the winter (Dalquest, 1942; Angerbjörn and Flux, 1995; Nunome *et al.*, 2014). This provides opportunities to perform comparative studies between color changing and non-color changing populations of hares within a species allowing to control for the genetic background variation and to determine the genetic basis of the phenotype (Dalziel *et al.*, 2009). The attempts to understand the bases of seasonal coat color change variation in hares are restricted to the work of Nunome *et al* (2014). However, no differentiation between populations of white winter morph and brown winter morph of *L. brachyurus* was found in three candidate genes related with coat color in mammals (*ASIP*, *TYR* and *MC1R*).

1.4. Snowshoe hares and seasonal coat color change

The snowshoe hare (*L. americanus*) is one of the species that undergo seasonal coat color change and will be used as model in this work. It is a widespread species in North America (Figure 1.2) and inhabits mainly forest areas (Flux and Angermann, 1990).

Recently, it has been proposed that snowshoe hares are divided in three Evolutionary Significant Units (ESU) – Boreal, Pacific Northwest and United States Rockies – based on the species genetic structure (Cheng *et al.*, 2014). Populations from the southernmost range show signs of genetic fragmentation but are the most unique genetically. Furthermore, evidences for mitochondrial DNA introgression from *L. californicus*, the black-tailed jackrabbit that has a brown winter coat, into the Pacific Northwest population of *L. americanus* have been described (Seixas, 2012; Cheng *et al.*, 2014; Melo-Ferreira *et al.*, 2014).

Snowshoe hares require areas with abundant understory vegetation, such as high grasses and shrubs, to use as forms to hide and rest during the day (Flux and Angermann, 1990; Hodges, 2000a). Unlike other hares, they do not build burrows to escape predation and do not flee until very nearly threatened and thus rely on their ability to stay still and camouflaged to avoid predators (Flux and Angermann, 1990; Zimova and Mills, 2014). Predation is in fact the major cause of death among hares (Hodges, 2000a, 2000b) and hares' behavior has been shown to be shaped by the presence of predators (Griffin *et al.*, 2005). For instance, hares' anti-predator behavior increases, *i.e*, tend to move less, in snowy winters with full moon when they are more likely to be seen by predators and when the predation risk is high (Griffin *et al.*, 2005). Therefore, they strongly depend on their coats for concealment while resting and foraging.

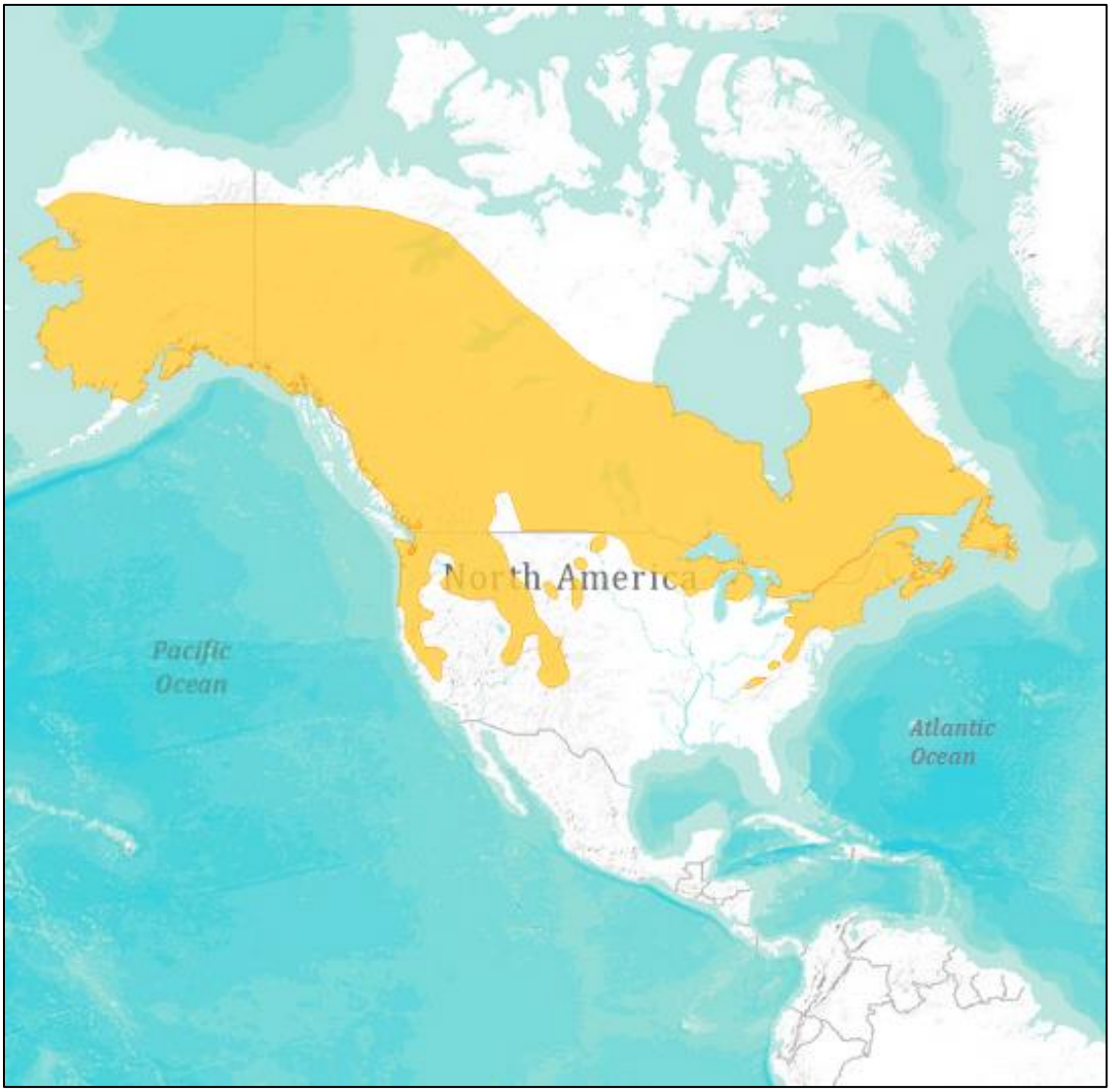

Figure 1.2 - The distribution of snowshoe hare (*Lepus americanus*) in North America. Adapted from IUCN (2008).

Snowshoe hares undergo two molts during the year, changing from a summer coat to a winter coat between October and November and changing back to the summer coat between April and May (Figure 1.3; Mills *et al*., 2013). Generally, the winter phenotype consists of a white coat color with black year tips. This white phenotype is produced by the guard hairs, since the underfur is tawny or pinkish, which determines that hares do not become completely white in winter, especially in the body extremities such as legs, feet, ears and face, where the hair becomes worn with use (Dalquest, 1942; Severaid, 1945). The summer phenotype is generally brown with a white belly and occasionally partly white hind feet that fail to lose the white winter hairs. The brown hairs of the summer coat are black-tipped, giving a variegated aspect to the upper body coat (Severaid, 1945). The

summer coat is also less bulky than the winter coat, and hairs grow flat against the skin (Dalquest, 1942). Little variation is reported between sexes (Severaid, 1945). According to Severaid (1945), the spring molt is less defined than the winter molt, being more difficult to define specific stages but generally in the spring the molt starts in the body and face in the direction of the feet and ears, while in the autumn molt the direction is reversed and starts in the ears and feet.

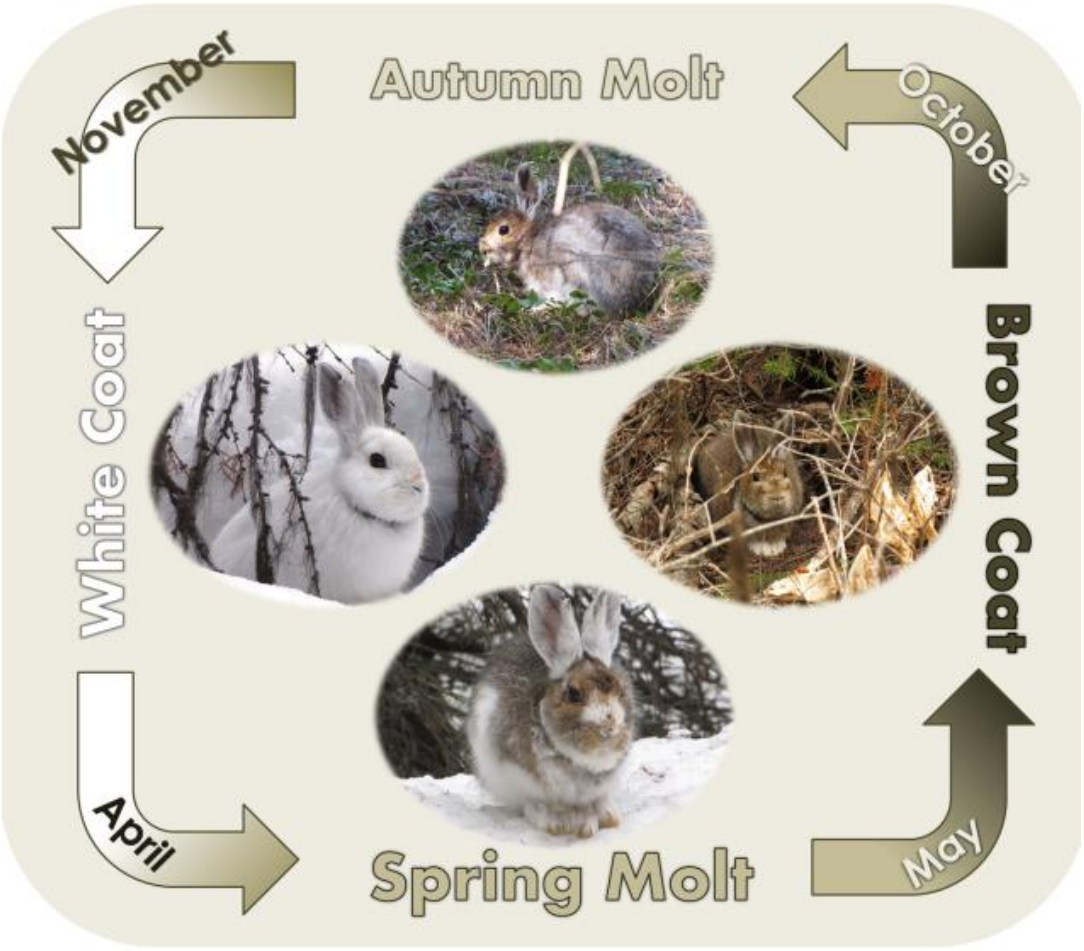

Figure 1.3 – The snowshoe hare molting cycle. Months of molting are based on Mills *et al.* (2013). Photographs are courtesy of L. Scott Mills.

Phenotypic plasticity of coat color phenology has been assessed recently at two study sites in Montana with different elevations and for a period of 3 years were variation in the snow cover duration was observed (Mills *et al.*, 2013; Zimova and Mills, 2014). Molt lasted approximately 40 days to be completed (Mills *et al.*, 2013; Zimova and Mills, 2014). In both

studies, initiation date was fixed both in the autumn and spring molt. However, although the rate of change of the autumn molt showed no phenotypic plasticity, the spring molt rate was not fixed and it was slower in the years of longer snow cover duration (Mills *et al.*, 2013). Nevertheless, the magnitude of the effect of predictors such as temperature or percent snow around hares was not enough to explain all the variation in the spring molt rate (Mills *et al.*, 2013; Zimova and Mills, 2014) suggesting that this factors are not the environmental cue that triggers the seasonal coat color change even if they can affect the rate of change. Furthermore, Zimova *et al*. (2014) found that hares do not modify their hiding behavior, flight response to an approaching predator or preference for a resting spot accordingly to the degree of color mismatch with the background. Despite the apparent lack of behavioral plasticity, it was reported that hares at the most elevated study site retained their winter coats longer than at the least elevated, suggesting a relationship between temperature and/or snow cover and molt duration (Zimova and Mills, 2014). Latitudinal differences have been also reported by Dalquest (1942), who described that generally the white phenotype exists in the northwest populations, while populations from Washington state varied from white to dark brown, with intermediate phenotypes being described as gray, light and dark tan. This is in accordance with current observations in Washington and Oregon States' populations where individuals that molt to winter phenotype and that retain the summer coat coexist. Interestingly, this is the population which has recently shown signs of introgression with *L. californicus* (Seixas, 2012; Cheng *et al.*, 2014; Melo-Ferreira *et al.*, 2014). The knowledge about the genetic mechanism behind the seasonal coat color change in this species is restricted to what has been referred before for other mammals.

1.5. Objectives

Considering that there is already some knowledge about the fitness consequences of the seasonal coat color change phenotype (Griffin *et al.*, 2005), its (lack of) plasticity (Mills *et al.*, 2013; Zimova and Mills, 2014) and a growing literature about the demography and phylogeny (Seixas, 2012; Cheng *et al.*, 2014; Melo-Ferreira *et al.*, 2014) of this species as well of other hares (Matthee *et al.*, 2004; Melo-Ferreira *et al.*, 2007, 2012; Alves and Hackländer, 2008), the snowshoe hare can be used as a valuable model to understand the genetic basis of seasonal coat color change, which would then allow studying the adaptive

relevance of the trait in a context of climate change. We hypothesize that the regulatory mechanisms that either block or allow the expression of the genetic pathway responsible for the production of melanin are involved in the seasonal expression of the coat color. Here, a gene expression study of the skin along the molt was performed to detect these regulatory changes and identify the machinery involved in seasonal coat color change.

The specific objectives of this work are:

- Using RNA-seq to sequence the skin transcriptome of five snowshoe hares undergoing spring molt and construct a *de novo* assembly of the skin transcriptome
- Following gene expression changes along the spring molting cycle, testing for differential expression of genes between time points in the molt that are defined by the color of the hairs growing in the skin – “white”, “molting” and “brown”.
- Identifying what genes and gene functions are involved in the molt.

This work is expected to provide valuable insights into the genetic determination of the seasonal coat color change, which can in the future be used to study the adaptive relevance of the trait in a molecular perspective and in a context of climate change.

2. Materials and methods

2.1. Sampling

For this work, samples were collected from 5 specimens undergoing spring molt captured from the Seeley Lake valley *Lepus americanus* population, in Montana, United States of America, in April 2012, and euthanized. This population is monomorphic regarding the coat color molt phenotype, since all individuals exhibit a white phenotype in winter and a brown phenotype in summer. For each individual, three skin biopsies where performed, each in a different part of the body displaying one of the three stages of molt – "white", "molting" or "brown". The stages were identified by the color of the hair growing in the skin. Therefore, each individual was a replicate, and sampling the three tissue types from the same individual allowed standardizing the information between tissue types. After collection, skin samples were immediately preserved in RNAlater and then stored at -80ºC until RNA extraction. In total, 15 samples were used, each representing the three different stages from the five different individuals.

2.2. RNA extraction, library preparation and sequencing

RNA extractions and library preparation for each sample had been performed in the University of Montana. Briefly, frozen skin samples were shaved and grinded in liquid nitrogen with the help of a ceramic mortar and pestle to make cell lysis more efficient. RNA extraction was performed using the RNeasy® Mini Kit from QIAGEN. After extraction, RNA quality and quantity was accessed using a NanoDrop IMPLEN P330 to calculate 260/280 and 260/230 ratios and using an Agilent's 2100 Bioanalyzer to calculate RNA concentration (µg/µL) and RNA Integrity Number (RIN). All samples had RIN values above 8, which is ideal for library preparation.

One µg of total RNA per sample was used to construct cDNA libraries for each sample using SureSelect Strand-Specific RNA Library Prep for Illumina Multiplexed Sequencing from Agilent Technologies. Briefly, olygo d(T) magnetic particles were used to capture the

poly(A) tails of mRNA contained in a sample. After this, the poly(A) RNA was enzymatically fragmented and the first strand of cDNA was synthesized. Afterwards, the second strand of cDNA was synthesized using deoxyuridine triphosphate (dUTP) instead of deoxythymidine triphosphate (dTTP) which marked it to be later degraded by uracil-DNA glycosylase (UDG), which allowed the production of a strand-specific library. After the formation of the second strand, end repair was performed and 3' ends of the cDNA were adenilated to allow adapter ligation. A PCR step was then used to amplify and index the adapter-ligated cDNA libraries. cDNA from each sample was individually barcoded which allowed multiplexing the samples in the sequencer while retaining their identity. During this step, UDG was introduced and the second cDNA strand was thus degraded. After concluding the library preparation, library size and the presence of adapter dimmer was inferred in an Agilent 2100 Bioanalyzer. KAPA Library quantification kit (KAPA BIOSYSTEMS) was used to estimate the molarity of the libraries with a qPCR and allow equal molarities of each library in the final pool.

All libraries were sequenced in two lanes of a HiSeq2000 and two lanes of a HiSeq 2500 (4 lanes in total) in the research facilities of University of California, Berkeley, producing 100bp paired-end sequence reads from the strand-specific libraries. All libraries were sequenced in each lane to avoid lane-specific biases.

2.3. Data and quality trimming

During sequencing, two of the lanes finished prematurely, leading to a truncation of the second read after the 70th nucleotide. Because of this, the CROP function of Trimmomatic (Bolger *et al.*, 2014) was used to discard the last 30 bases in these two lanes. After this, treatment of sequence data was performed equally for all reads from all sequencing lanes. The quality of the samples was checked with FastQC v0.10.1 (Andrews, 2014) and duplication levels assessed using custom scripts (Corsat, T., personal communication). The Illumina CASAVA-1.8 FASTQ Filter (Gordon, 2011) was used to filter out reads flagged by the CASAVA-1.8 as failing quality control. Then, technical sequences such as adapters were removed using Cutadapt 1.3 (Martin, 2011) and quality trimming was performed with Trimmomatic 0.32 (Bolger *et al.*, 2014). In the Trimmomatic command, it was specified that the first thirteen bases of every read would be cut due to a bias

observed in nucleotide composition that is possible due to the non-random hexamer primer activity of the primers used in library preparation (Hansen *et al.*, 2010). In addition, a 4bp trimming sliding window approach with a 15 phred quality score threshold was used. In the end, synchrony between read 1 and 2 was checked with *fastq_sync.pl* script (Johnson, 2013).

2.4. De novo assembly, annotation and assembly filtering

De novo assembling was performed using Trinity 2013-11-10 (Grabherr *et al.*, 2011) with the default parameters, including all obtained sequence data. *TrinityStats.pl*, a script included in the Trinity package, was used to assess the quality of the assembly by calculating general parameters such as number of reads assembled, N50 and mean contig length. Also, the Trinity algorithm has the capacity to predict putative genes and isoforms, which it classifies as components and isoforms, respectively (Grabherr *et al.*, 2011). Therefore, *TrinityStats.pl* also reported the number of components and isoforms in the assembly. Contig annotation was performed using the rabbit (*Oryctolagus cuniculus*) peptome present in the ENSEMBL database (Flicek *et al.*, 2014), downloaded using BioMart (Kasprzyk, 2011). Blastx, implemented in Blast+ 2.2.29 (Camacho *et al.*, 2009), was used to blast the transcriptome against the reference, requiring a maximum e-value of 1e-20 and only reporting the best database target for each contig. The script *analyze_blastPlus_topHit_coverage.pl* of the Trinity package was used to calculate the percentage of length covered or completeness, defined as the percent length of the database covered by the transcriptome. For each blast hit the percentage of the database protein covered by the contig was calculated.

During the *de novo* assembly it is likely that some contigs are produced erroneously, meaning that they are chimeras, resulting from the assembly of two or more reference genes, or dubious predicted contigs that originated from lowly covered and/or highly similar regions, such as alleles, isoforms or paralogous genes that result in the formation of redundant contigs (Cahais *et al.*, 2012; Singhal, 2013). In fact, mistakes such as formation of chimera and report of dubious predicted isoforms have been observed in Trinity produced assemblies (Vijay *et al.*, 2013). Also, isoform prediction is specially untrustworthy in a *de novo* assembly for lowly expressed isoforms (Angelini *et al.*, 2014).

In an effort to mitigate this problem, contigs that were not annotated during the blastx were excluded from the assembly. Although this might also exclude correctly inferred contigs/genes that are not annotated in the reference database, we used this conservative approach to increase the probability to include only real genes in the final assembly.

2.5. Mapping and relative abundance estimation

Relative abundances were calculated using RSEM 1.2.8 (Li and Dewey, 2011), which is specially designed to perform abundance estimations with reads mapping to a transcript set rather than a reference genome (Li and Dewey, 2011). This allows using a *de novo* transcriptome, as the one constructed here. Furthermore, it has the advantage of dealing with reads aligning to multiple regions in the transcriptome rather than excluding them from the analysis, as other approaches do. This program uses Bowtie 1.0.0 (Langmead *et al.*, 2009) to align reads back to the transcriptome, reporting as valid alignments those with less than 2 mismatches in the first 23bp. RSEM's algorithm then determines the likelihood of the alignment (Li and Dewey, 2011). RSEM was run with default parameters whereas Bowtie was used establishing a 23bp seed length (--seed-length) and allocating 512mb of memory for the calculation of the best first alignment (--bowtie-chunkmbs). RSEM has also the advantage of outputting relative abundances based on the Trinity components and isoforms. Beyond this point, the analysis was performed with abundance estimates for the components.

2.6. Differential gene expression analysis

Differential expression analysis between the three different molt stages was evaluated using edgeR (Robinson *et al.*, 2010) in RStudio v0.98.976 using R 3.0.3 (R Core Team, 2014). edgeR applies a Generalized Linear Model (GLM) to accommodate an experimental design with a non-normally distributed response variable, which is the case of count data. The GLM implemented in edgeR uses a negative binomial (NB) distribution to model read counts (McCarthy *et al.*, 2012). Although read count data is often fitted to a Poisson model, this has been proven to be too simplistic for RNA-seq data because it

assumes less variation than it actually exists due to biological variation between samples (Robinson and Smyth, 2007). If Y_{gi} symbolizes the read counts for sample *i* and gene *g*, the NB distribution is fitted as,

$$Y_{gi} \sim NB(\mu_{gi}, \varphi_g)$$

Where the mean (μ_{gi}) can be obtained by multiplying the library size with the relative abundance of the gene in the sample and the dispersion (φ_g), as demonstrated by McCarthy and colleges (2012), is

$$var(Y_{gi}) = \mu_{gi} + \varphi_g \mu_{gi}^2$$

which dividing by μ_{gi}^2 becomes:

$$CV^2(Y_{gi}) = 1/\mu_{gi} + \varphi_g$$

The first term represents the squared technical coefficient of variation (TCV) and the second the squared biological coefficient of variation (BCV). edgeR allows the estimation with a Cox-Reid adjusted profile likelihood assuming that all genes have the same dispersion (common dispersion), and that the dispersion follows a trend (trended dispersion) or calculating a dispersion value for each gene (tagwise dispersion) (McCarthy *et al.*, 2012). The common and trended dispersion are most often oversimplifications since each gene has most likely its own dispersion (McCarthy *et al.*, 2012). However, if the number of samples is small, the estimation of BCV will not be representative of the real biological variability, skewing the tagwise dispersion. Because of this, Robison and Smith (2007) implemented a weighted likelihood empirical Bayes approach that uses the common dispersion (that since it is calculated from all the genes is likely more representative) to improve the tagwise dispersion estimates for each gene, and this is the approach implemented in edgeR. QQplots of goodness of fit statistics can be drawn to understand how these three different dispersions fit the data (McCarthy *et al.*, 2012).

Steps in the edgeR analysis consisted first of excluding components that did not have at least 1 count per million (CPM) mapped reads (which represents roughly 8-19 mapped reads to a contig) in at least 5 samples. This can be considered as another filtering step that will exclude data from misassembled contigs. It has been shown that the lowest expressed transcripts are very often False Positives (FP), *i.e.*, transcripts that are wrongly predicted to be expressed (Angelini *et al.*, 2014). Angelini and colleges (2014) have shown

that filtering these lowly expressed transcripts increases the precision – fraction of True Positives (TP) relatively to the total (TP+FP) – of methods such as RSEM. Furthermore, it has been shown that filtering for low expression/coverage eliminates chimera and redundant or dubious contigs, that might have resulted from assembly errors (see above; Cahais *et al.*, 2012; Yang and Smith, 2013).

Second, to allow the comparison of expression estimates across samples (libraries) and features, a normalization step was performed. The trimmed mean of M-Values (TMM) is the method implemented in edgeR to normalize data across libraries. The TMM takes into account the underlying differences of the RNA population between samples and calculates a scaling factor to be applied to the read counts (Robinson and Oshlack, 2010). According to Robinson and Oshlack (2010), estimates of expression level will vary not only with library size or gene length (the greater the length, more likely that a gene will have more reads mapped to it) but also with the population of genes being expressed in each sample. For example, consider two samples A and B, where a given set of genes is expressed in both samples. However, sample A has a particular group of genes expressed that are not expressed in B, meaning that sample A has twice the number of genes being expressed comparing to sample B. If the sequencing effort is equal for both samples, the expression level of a gene that is both expressed in A and B will be the double in B than it is in sample A. The TMM normalization assumes that the majority of genes is not differentially expressed and therefore will search for these genes to calculate a scaling factor between samples. By choosing one reference sample, the algorithm will calculate the gene-wise log-fold change (M values) between samples and the reference and the absolute expression levels (A values) in each sample. The principle here is that after trimming x% of the M and A values (*i.e.*, genes with high log fold changes or high read counts), the genes that remain are those not differentially expressed and should have the same expression value across samples. Therefore, they are suitable to be used to calculate the scaling factor to normalize the samples expression levels. The scaling factor for each sample is the weighted mean of M values, after trimming.

Third, the common, trended and tagwise dispersion were calculated using a Cox-Reid adjusted profile likelihood. The GLM was then fitted to the data using a log-linear model to link the response variable (count data) and the predictors. The predictors used were “Molt Stage” and “Individual” as a blocking factor. The blocking factor is used to group samples according to their individual origin, controlling for underlying variation between individuals

that might influence the results and make the variation between molt stages stand out (Quinn and Keough, 2002). A likelihood ratio test (LRT) was used to test for differential expression in three pairwise comparisons ("white" vs "brown", "white" vs "molting" and "molting" vs "brown") and a Benjamin-Hochberg multiple test correction was applied. Genes with a false discovery rate (FDR) lower than 0.05 were identified as being differentially expressed.

To better interpret expression patterns, a hierarchical clustering analysis was performed to group genes and samples by their expression patterns. The analysis was done using expression levels calculated as fragments per kilobase of exon per million reads mapped (FPKM) using read counts normalized as stated above, log2 transformed and mean-centered to reduce the effect of outliers. The correlation of expression levels between samples was calculated using a Pearson's correlation coefficient to understand how molt stages correlate. Expression levels of genes that were differentially expressed in at least one comparison were used to calculate the Euclidian distance. Then, a complete linkage hierarchical clustering analysis was performed to draw the patterns of similarity of correlation coefficients between samples and of the genes expression levels. The package ggplots 3.4.1 (Warnes *et al.*, 2014) was used to produce the heatmap.

Using the same expression levels (log2 FPKM, mean centered) a partitioning type clustering analysis was performed to encounter groups of genes with similar expression patterns. Because a k number of clusters needs to be assessed for this analysis, two methods were used to estimate k for comparison. First, a gap statistics approach was used to infer the number of k clusters, as implemented in the R package cluster v1.15.2 (Maechler *et al.*, 2012). This methodology compares the change in the within cluster dispersion with the expectation under a null reference distribution that is generated from sampling uniformly from the original population with a Monte Carlo algorithm (Tibshirani *et al.*, 2001). The function clusGap from the package referred above was used establishing 1000 bootstraps resampling and the PAM clustering algorithm. Then, silhouettes were also used to estimate the number of k. The rationale consists on performing the clustering with several values of k and then to determine which k maximizes the average silhouette width. The silhouette value s(*i*) compares the average distance between a given object with all the others that are inside the same cluster C_i and the smaller distance between the same object and another outside the cluster (Reynolds *et al.*, 2006). The mean of these values for all the objects in a cluster is the average silhouette width and the best cluster is then

the one that maximizes this value. Therefore, the PAM function of the package cluster was used to perform the partition clustering analysis using the Partitioning Around Medoids (PAM). Traditionally the kmeans method, which uses the means of the points in the cluster as coordinates to form centroids and looks for minimizing the sum of square distance between points, has been used to cluster gene expression data (D'haeseleer, 2005). PAM is a generalization of kmeans that instead of centering data around centroids centers around medoids that aim at reducing the dissimilarity within the clusters (Reynolds *et al.*, 2006). Here, we repeated the analysis with k ranging from 2 to 20 in order to access the k that maximized the average silhouette width. Because different values of k were obtained with the Gap statistics (k=4) and the silhouette method (k=3), results with k=3 and k=4 were produced for comparison. Finally, using the *plot_expression_patterns.pl* script provided by the Trinity pipeline, the expression values of clusters of genes with common expressed profiles were plotted.

2.7. Gene-Ontology term enrichment analysis

To understand whether differentially expressed (DE) genes were enriched in particular functions, a Gene Ontology (GO) term enrichment analysis was performed. As explained by Bauer *et al.* (2008), the principle of this analysis is to annotate with GO terms the population of genes (in this case the transcriptome) and the groups of genes of interest, such as the groups of genes differentially expressed in the three comparisons made in the differential expression analysis. Then, by sampling at random a number of genes equal to one of the study sets from the population, the probability of drawing the same or higher number of genes annotated to a given GO term is calculated. If that probability is low (and smaller than a given threshold) the GO term is significantly enriched in the study set.

A custom annotation was performed for the rabbit (*Oryctolagus cuniculus*) with the help of BioMart tool (Kasprzyk, 2011) at the ENSEMBL (Flicek *et al.*, 2014) website and custom scripts. Then this annotation was used to make a GO analysis in Ontologizer 2.1 (Bauer *et al.*, 2008). The implementation of the Parent-Child test is an advantage in Ontologizer since this method takes into account the true-path rule that results from the hierarchical nature of the GO term annotation and implies that when a gene is annotated to a GO term it will also be annotated to all the parents of that term which will generate overlapping

annotations (Grossmann *et al.*, 2007). Therefore the Parent-Child-Union test was used, with the Benjamini-Hochberg multiple test correction. Significance was established at a FDR<0.05.

2.8. Candidate gene perspective

Using the ENSEMBL annotations and associated gene names for each differentially expressed component, a final search in the list of differentially expressed genes was performed for gene names related with melanin synthesis pathway, melanosome transporters and circadian rhythm related genes based on literature on the subject (Bennett and Lamoreux, 2003; Slominski *et al.*, 2004; Meredith *et al.*, 2006; Hoekstra, 2006; Lin *et al.*, 2009).

3. Results

3.1. Raw data and quality control

After library preparation, the 15 strand-specific cDNA libraries, with fragment length varying between 252 and 363bps (including adapters), were paired-end sequenced in 2 lanes of a HiSeq 2000 and 2 lanes of a HiSeq 2500, totaling 4 lanes, producing a total of 385887081 100bp read pairs. After the quality filtering steps, which removed adapters and low quality reads, the number of read pairs was 323917483, with sizes ranging between 23 and 87bp. Percent of unique inserts ranged between 89 and 97%. All cleaned reads were assembled *de novo* using Trinity.

3.2. Assembly and annotation

The assembling yielded 321441 contigs (representing 239909 components) in the raw assembly (Table 3.1). Filtering out the contigs with no Blastx hit, *i.e.* with no annotation, the number of contigs was reduced to 85967, representing 40001 Trinity components. The 85967 contigs blasted against 15105 proteins or 14421 genes (Table 3.1). Contig lengths varied between 201bp (as Trinity, by default, excludes all contigs smaller than 200bp) and 33280 bp, before and after filtering.

Table 3.1 – Summary of assembly parameters before and after annotation against the rabbit peptide reference

	Raw assembly	**Annotated assembly**
Nr. assembled bases	287,693,282	176,947,733
Nr. Trinity componentes	239909	40001
Nr. Trinity isoforms/contigs	321441	85967
Contig N50	2087	3360
Median contig length	376	1504
GC contente (%)	52.97	53.72
Nr. of gene annotations		14421
Nr. of protein annotations		15105

Completeness, defined as the percent length coverage of the reference proteins by the contigs in the assembly (Martin and Wang, 2011), was calculated and is shown on Figure 3.1. Because each reference protein is, for some cases, the best Blastx hit for more than one contig, only the contig that provided the highest coverage was used to draw the histogram. From that it is possible to see that 11162 of 15105 (74%) of the reference proteins were covered in more than 70% of their length by the transcriptome.

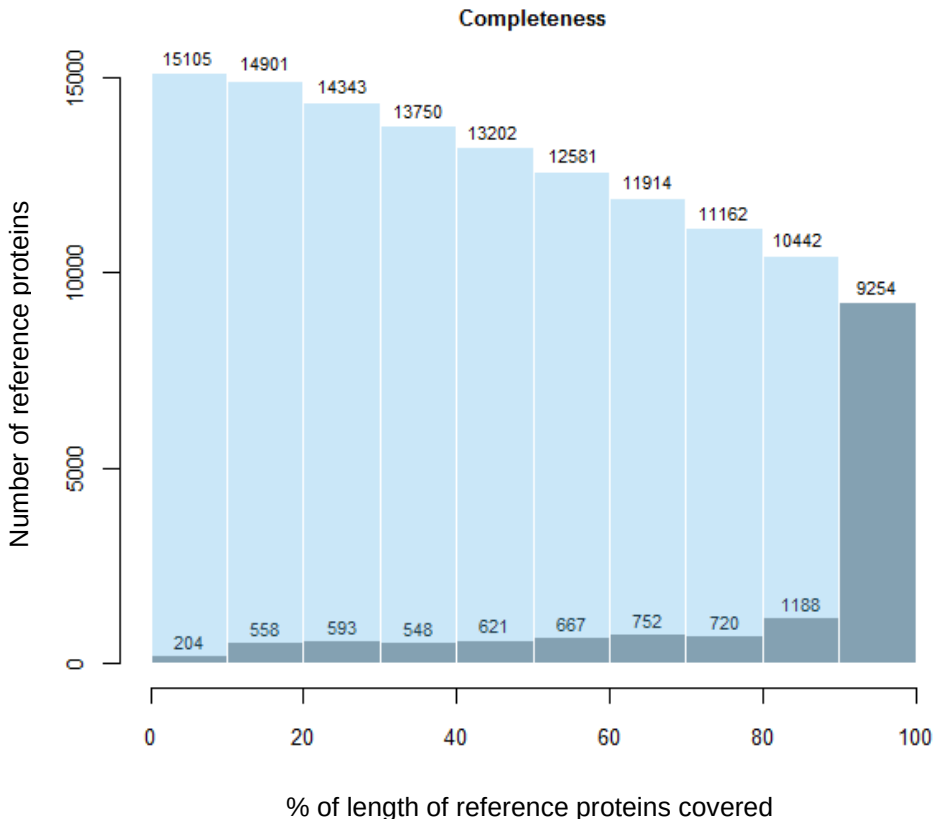

Figure 3.1 – Histogram of frequency (gray) and cumulative frequency (blue) of the percent length coverage (completeness) of the reference proteins by the contigs in the assembly. Bar height represents the number of reference proteins covered in the given percentage class.

3.3. Mapping and relative abundance estimation

Mapping reads back to the transcriptome using bowtie, as implemented by RSEM, resulted in a mapping percentage, here defined as the proportion of reads that successfully mapped back to the transcriptome, of approximately 73%. Since RSEM generates abundance estimates for both Trinity components and isoforms, and it refers to them as “gene” and “transcript” respectively, the results here reported are for a gene level analysis. For a matter of simplicity and for a first approach, this work focuses on the gene results generated by RSEM. Therefore, the term “gene” or “genes” will for convenience be used as synonymous of “Trinity component(s)”.

3.4. Differential gene expression analysis

3.4.1. Filtering lowly expressed components

The initial step of the differential expression analysis, which consisted in only keeping contigs with counts per million (CPM) mapped reads higher than 1 in at least five or more samples reduced the number of putative genes (Trinity components) to 14568 that blasted against 12210 unique ENSEMBL gene codes. Then, in order to understand how different factors, such as assay design, might have affected the quantitative measures of gene expression, read files were maintained separated and counts were also obtained per lane.

3.4.2. Assessment of variability and dispersion in the dataset

Counts were plotted in a multidimensional scale where the distance between samples was calculated as the average root-mean-square (Euclidian distance) of largest absolute log2 fold change between each sample (Figure 3.2). This showed that the logFC for each lane for the same individual tended to overlap, suggesting that variation between lanes is small

and therefore the results of each sample could be collapsed (Figure 3.2a). Nevertheless, although there is an aggregation of "molting stages" observable on dimension 2, with "white" on top, "brown" on the bottom and "molting" between both, the first dimension separates individual ame12 from all the others, suggesting that variation between individuals might influence the results and thus should be a factor to take into account in the statistical analysis. Because of this, in the differential expression analysis and GLM model counts were grouped by individual to block any differences between them (see Materials and Methods). After concluding this step, the results of the different lanes were concatenated (the different fastqc files were concatenated and RSEM was run again to produce new abundance estimates) and the results observed before - clustering by individual and "molting" stage - remained (Figure 3.2b).

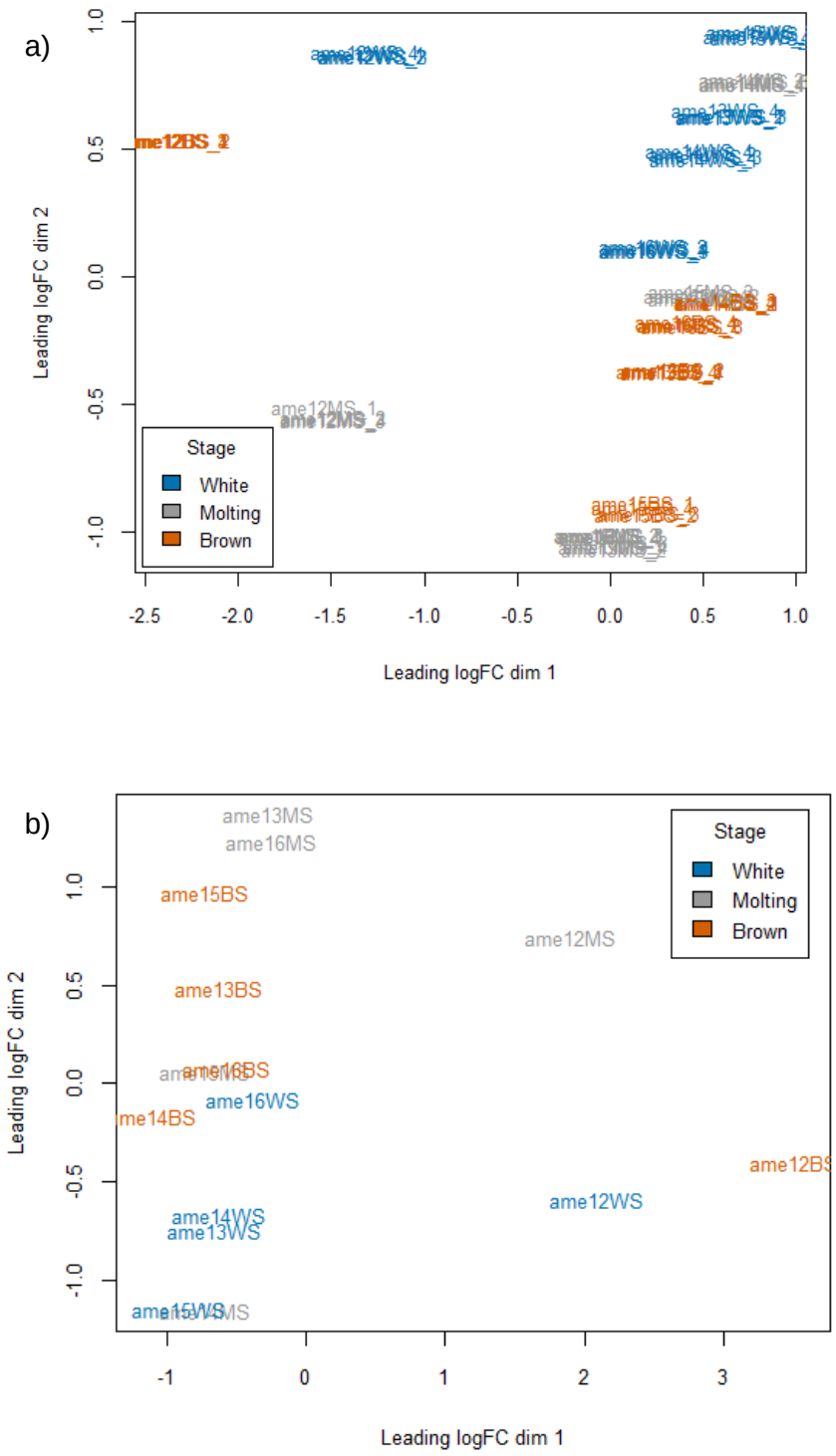

Figure 3.2 - Multidimensional scaling plot of the distance between samples according to their leading log2 Fold Change. In a) expression estimates for each lane were kept separate in order to understand the overlap of each sequencing replicate and in b) expression estimates were obtained by concatenating results for each lane and therefore only an unique point for each individual is seen. Labels of the points represent the individual code of each sample with the "molting" stage symbolized as WS ("white"), MS ("molting"), BS ("brown"), plus a number that identifies the lane in plot a). Points are colored according to the molt stage.

Dispersion estimates assuming a common, trended or tagwise dispersion were calculated in edgeR (Figure 3.3). The common dispersion estimated a biological coefficient of variation (BCV) of 22%, meaning that the abundance for each gene between biological replicates varies 22% (McCarthy *et al.*, 2012).

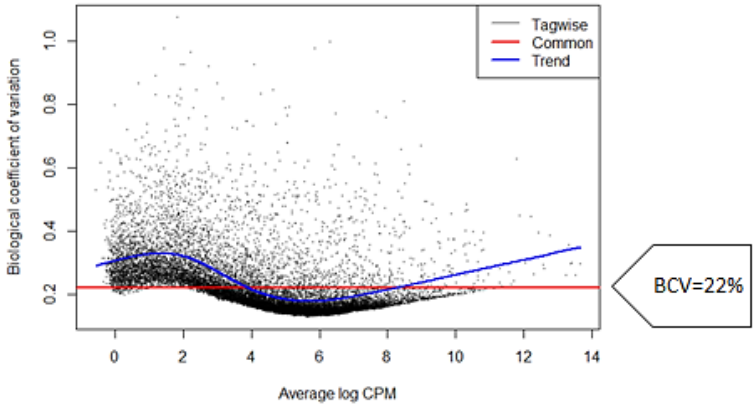

Figure 3.3 - Plot of biological coefficient of variation (BCV), calculated either assuming a common value of dispersion (red), trended value (blue) or tagwise (black), as a function of the average log counts per million (CPM) of mapped reads.

QQ-plots were used to assess the goodness of fit of the different dispersion allowed in edgeR, which suggested that the common and the trended dispersion do not provide a good fit to the dataset since a group of genes show greater variability than the one assumed by the first two methods (Figure 3.4, in blue). This is a statistical justification for the use of a tagwise method to calculate the biological coefficient of variability.

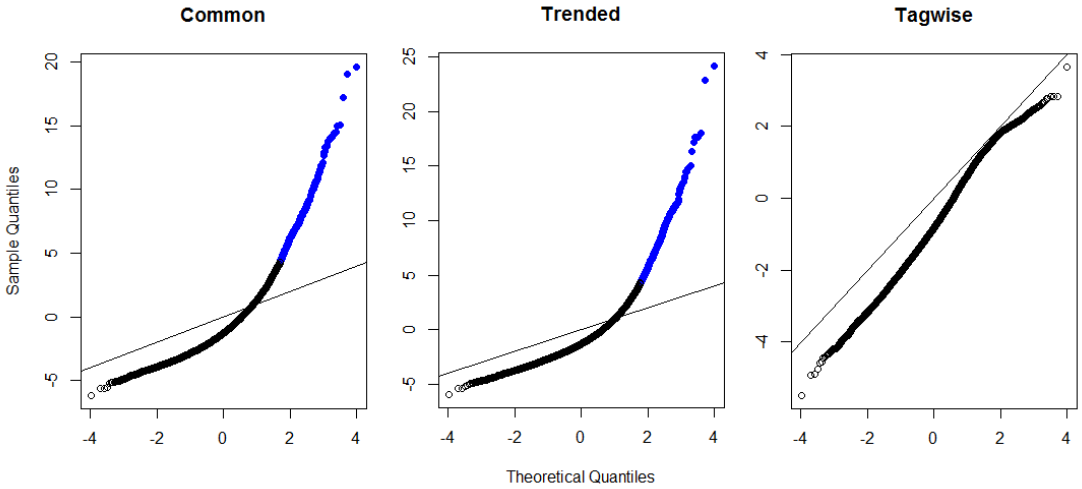

Figure 3.4 - QQ-plots showing the fitness of common, trended and tagwise dispersions to the data. Genes that show a significant poor fit (Holm-adjusted P-value < 0.05) are shown in blue.

3.4.3. Number of differentially expressed genes

Differential expression analysis was carried on next and, overall, of the 14568 genes, 943 were found to be differentially expressed (DE) (FDR < 0.05) in at least 1 comparison between stages (Figure 3.5). The complete list of differentially expressed genes in each pairwise comparison as well as their annotations is provided in Appendix 1 – Table S1. Of this, only 10 genes were common to the three comparisons and more than half were exclusively differentially expressed between "brown" and "white" (568; approximately 60% of all DE genes). The highest number of DE genes was found for comparisons with the "white" stage, with 739 and 312 genes being DE between "white" and "brown" and "white" and "molting", respectively (Table 3.2,Figure 3.5 and Figure 3.6).

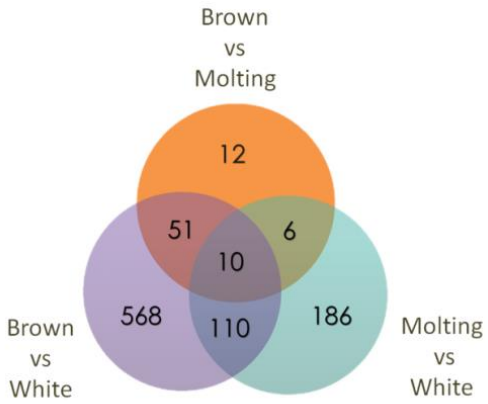

Figure 3.5 - Venn diagram showing the overlap of differentially expressed genes between comparisons of "molting" stage.

There was a higher number of up regulated genes in the "white" stage (Table 3.2 and Figure 3.6). It was also noticeable that fewer genes were found to be DE between "brown" and the "molting" stage, relatively to the other two comparisons.

Table 3.2 – Number of differentially expressed genes (FDR<0.05). For each pairwise comparison results are given relatively to the second referred molt stage. For example, there are 480 genes upregulated in "white" relatively to "brown". If a gene is differentially expressed, upregulation is determined when the log fold change is positive and downregulation when fold change is negative (see Figure 3.6).

Comparison (GLM paired analysis)	#Differentially expressed genes	Upregulated	Downregulated	Not differently expressed
Molting vs Brown	79	39	40	14489
White vs Brown	739	480	259	13829
White vs Molting	312	279	33	14256

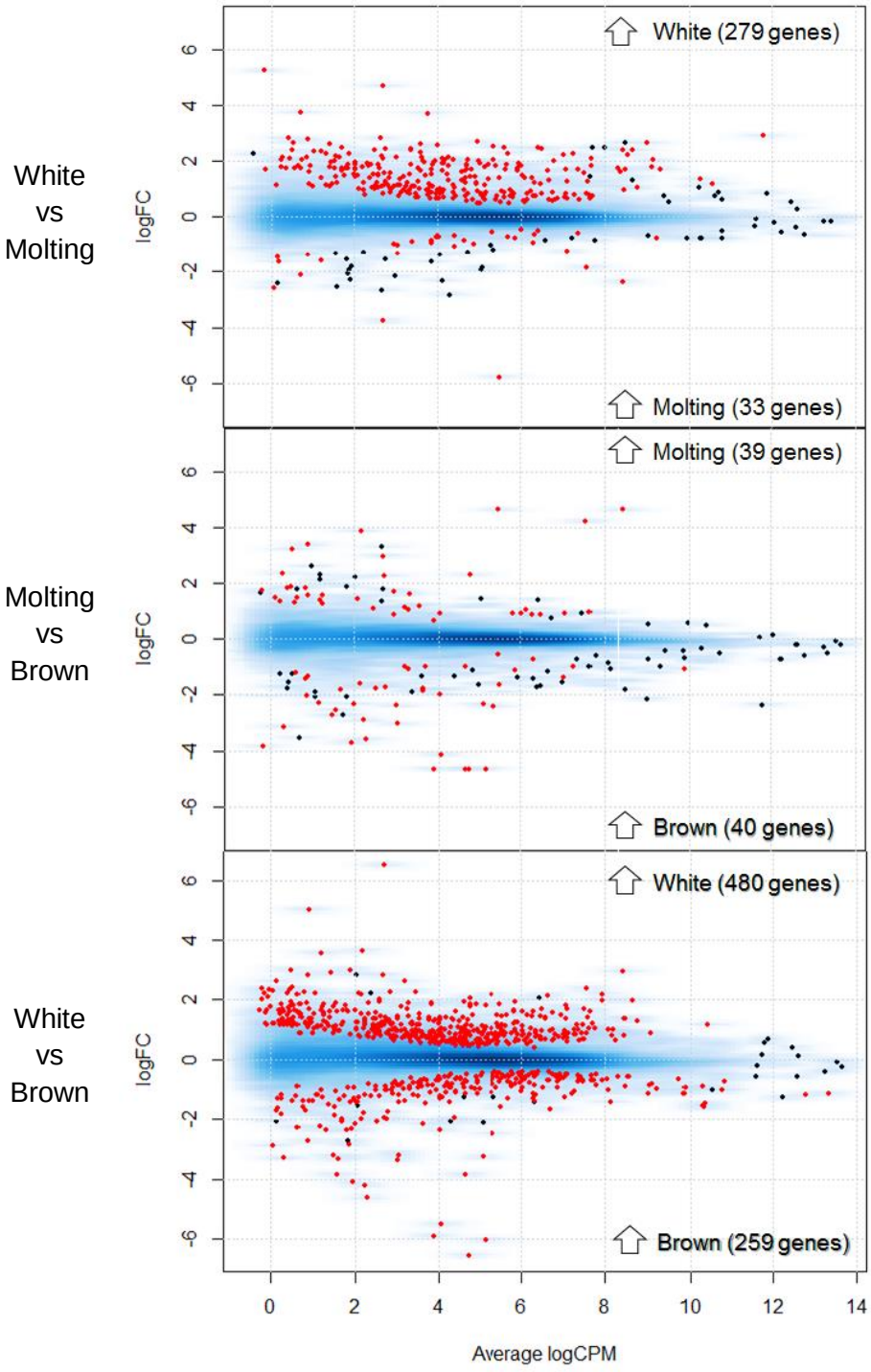

Figure 3.6 - Smear Plots of the average log counts per million (CPM) reads plotted against log Fold Change. For each of the three comparisons made, differentially expressed genes (FDR<0.05) between the molt stages referred in each graph are shown in red. Arrows symbolize upregulation in the respective molt stage of the number of genes in brackets.

3.4.4. Gene expression patterns

Hierarchical clustering of gene expression levels resulted in the formation of four clusters, with samples being grouped according to molting stage (Figure 3.7). “Molting” stage samples were divided between the two main clusters, which separated “white” and “brown” stage samples. It is discerned a change in the global expression pattern as the molt progresses from “white” to “brown” and “white” stage shows the highest levels of expression.

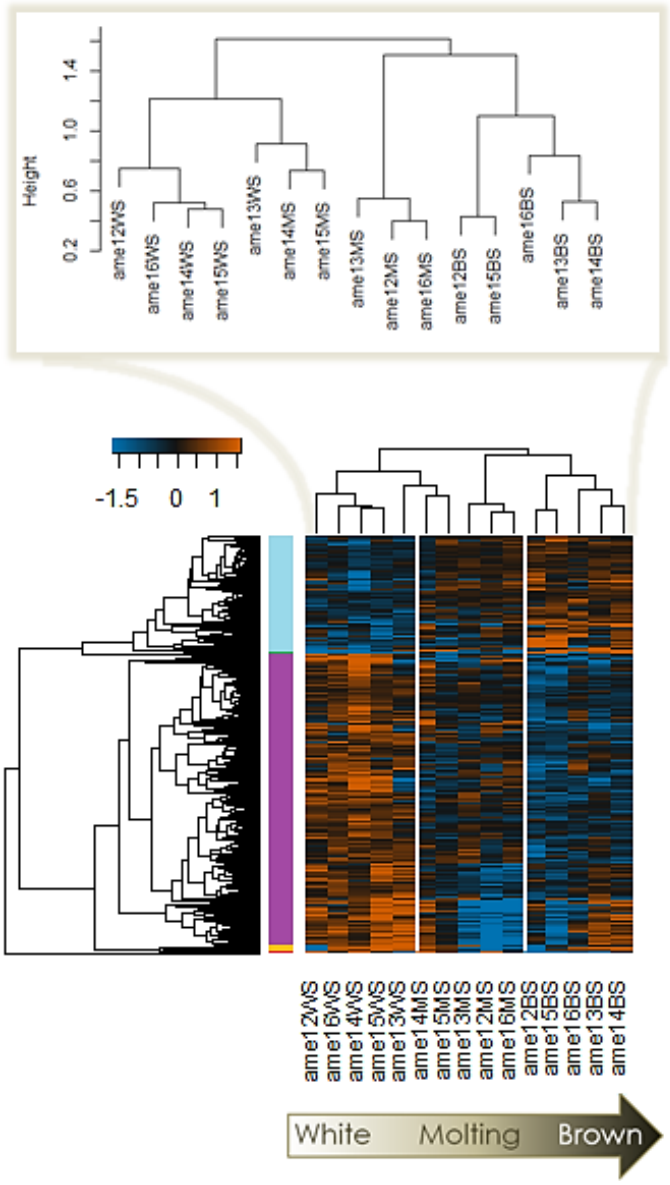

Figure 3.7 - On the top, a dendrogram of samples clustering with correlation of distances between nodes, depicted on the yy-axis. On the bottom, the heatmap of log 2 transformed, mean centered FPKM levels of expression per DE gene (row) and sample (column). The dendrograms are resultant of complete linkage hierarchical clustering of Pearson’s correlation coefficient between samples and Euclidian distances between gene expression levels, which is displayed in the yy axis.

To better understand the patterns of expression, partitioning clustering of expression was performed to extract qualitative clusters of genes with similar expression patterns. Prior to the analysis, as explained in the Materials and Methods section, the number of k was estimated with two approaches, the Gap statistics and the silhouette method. While the Gap statistics estimated k=4, the silhouette method estimated k=3 (Figure 3.8).

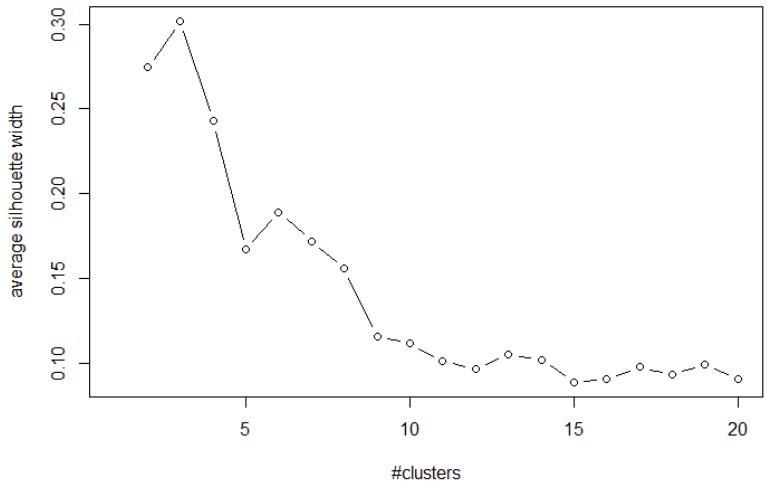

Figure 3.8 - Average silhouette with of the centered log 2 FPKM expression levels for PAM clustering of k=2:20. The maximum average silhouette width is k=3.

Clustering was thus performed with both values of k and at least three distinct expression patterns stand out from the clusters: 1) an overall increase in expression as the molt progresses (cluster 2 in Figure 3.9 and 3.10), 2) a decrease in expression along the molt (cluster 3 in Figure 3.9 and cluster 3 and 4 in Figure 3.10), and 3) a decrease in expression in the intermediate stage of the molt compared to the other two (cluster 1 in Figure 3.9 and 3.10). It is noticeable that this last pattern was not consistent in all individuals or genes since individual ame14 and ame15 follow either a stable or a decreasing change in expression, respectively. A decrease in expression along the molt is the most frequent pattern in the dataset (either 507 in k=3 or 516 in k=4), followed by an increase and an intermediate pattern. The complete list of Trinity components in each cluster is provided in Appendix 1 – Table S2. The classification made for the expression patterns is merely qualitative and because the clustering with k=4 does not add a new expression pattern to those found with k=3, the clustering results from k=3 were used as reference for further analyses.

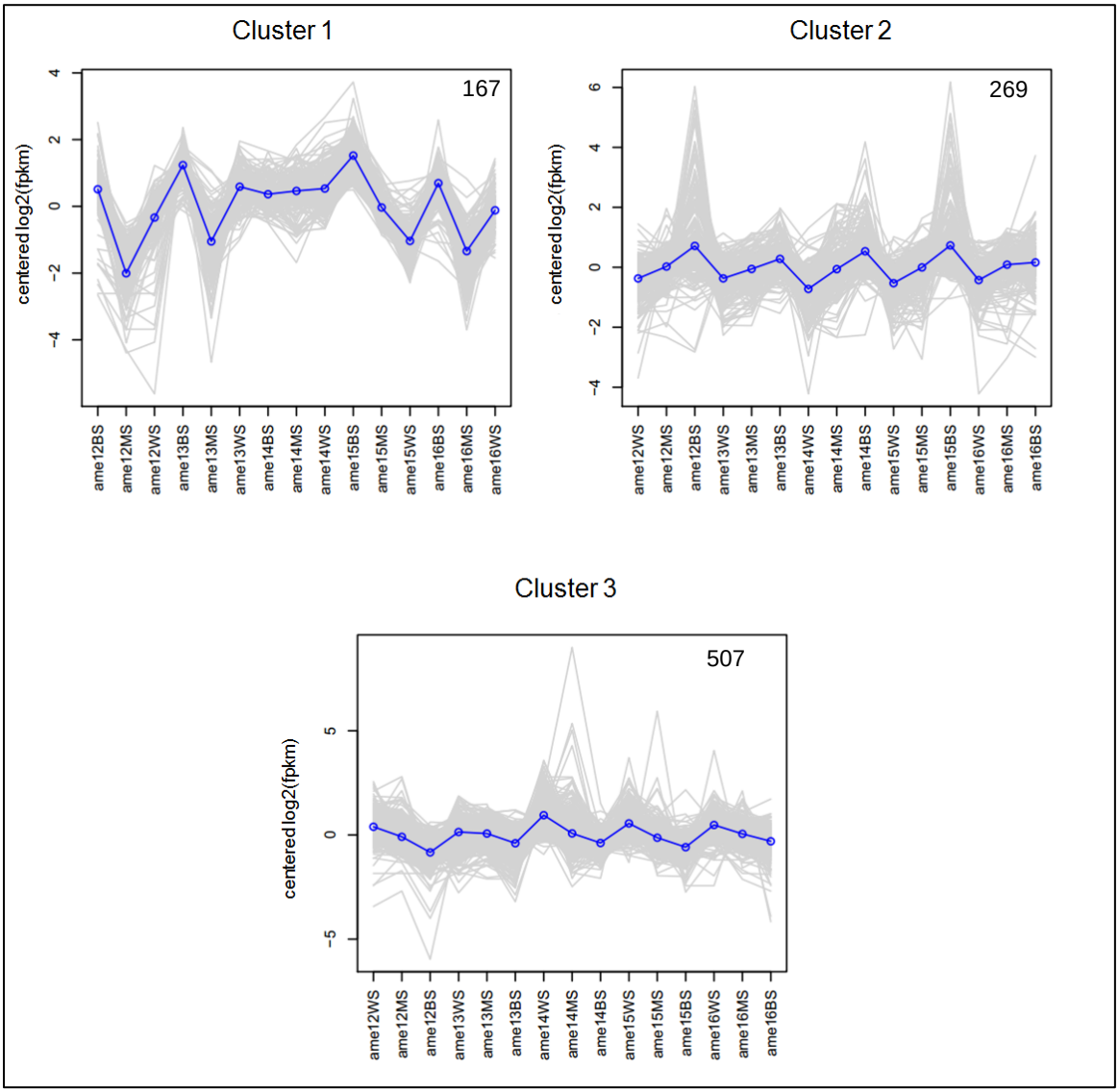

Figure 3.9 – Gene clusters obtained from PAM clustering analysis with centered log2 FPKM levels of expression for each gene establishing k=3. Numbers at the right top corner of each graph symbolize the number of genes in each cluster.

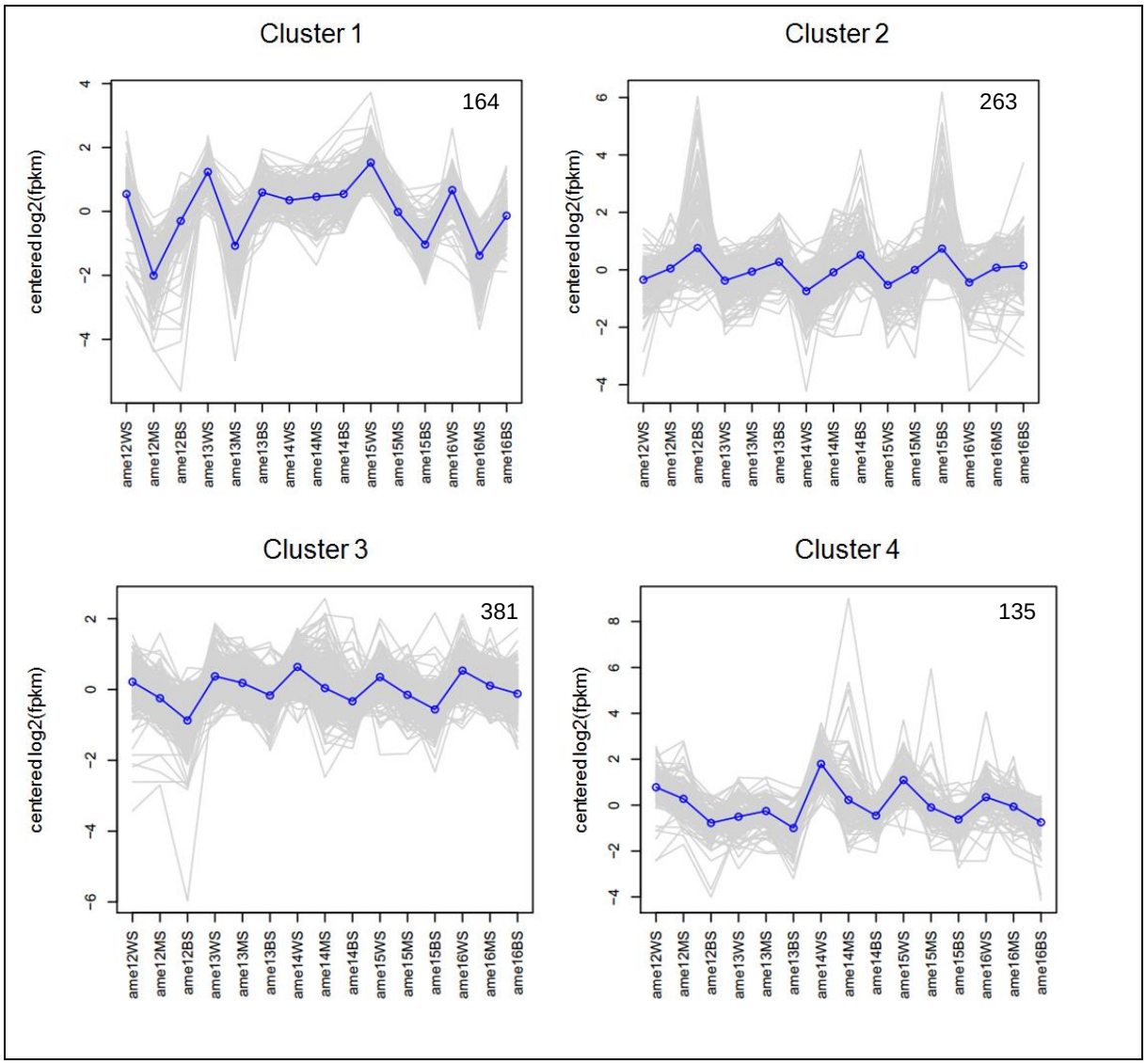

Figure 3.10 - Gene clusters obtained from PAM clustering analysis with centered log2 FPKM levels of expression for each gene establishing k=4. Numbers at the right top corner of each graph symbolize the number of genes in each cluster.

3.5. Gene-Ontology term enrichment analysis

To identify GO categories significantly enriched in our dataset, a GO enrichment analysis was performed. From the population of 14421 gene annotations, it was possible to annotate 12912 with Gene Ontology terms in all of the three main categories. Of notice is the fact that each Trinity component is composed of several sequences (*i.e.*, isoforms) that in a few occasions have its best Blastx hit in different genes. Because of this, the number

of ENSEMBL and GO annotations is not the same as the number of DE genes (*i.e*, Trinity components) (Table 3.3). In this analysis, all the multiple ENSEMBL codes were used to produce the GO term annotation.

Table 3.3 – List of the number of differentially expressed genes (i.e Trinity components), the number of annotations to ENSEMBL genes and number of GO terms associated with the ENSEMBL gene codes for each pairwise comparison between the different "molting" stages.

Pairwise Comparison	# DE Trinity Components	#ENSEMBL Gene Codes	# ENSEMBL codes annotate to GO terms
Molting vs Brown	79	93	91
White vs Molting	312	324	308
White vs Brown	739	733	676

The results of the enrichment analysis, for each of the GO term categories and differentially expressed genes in each pairwise comparision, are displayed in Figures 3.11 to 3.14 and in Appendix 1 – Table S3 to S5. The lower number of significant enriched terms is encountered among the differentially expressed genes between "molting" and "brown" stage (Figure 3.11) with only one term per GO category, which might be related with the fact that this is the pairwise comparison showing the smaller number of DE genes (Table 3.2).

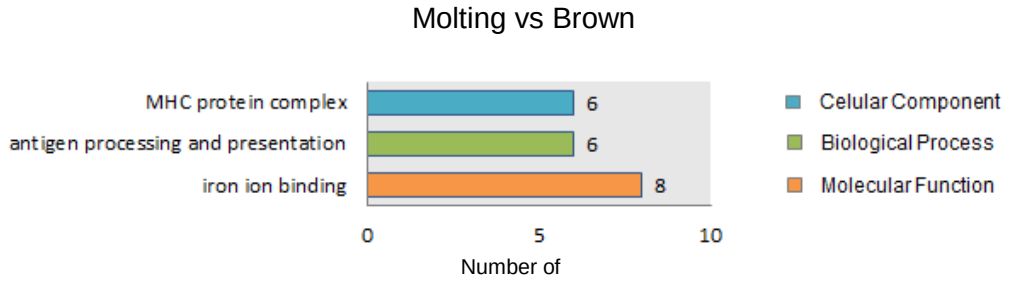

Figure 3.11 - Gene Ontology enriched terms (FDR<0.05) of all GO categories for genes being differentially expressed between "molting" and "brown" stage. The different colors of the bars represent the three different GO categories. Numbers next to the bars in the graphs represent the number of genes (in this case, ENSEMBL gene codes) that are associated with the respective GO term.

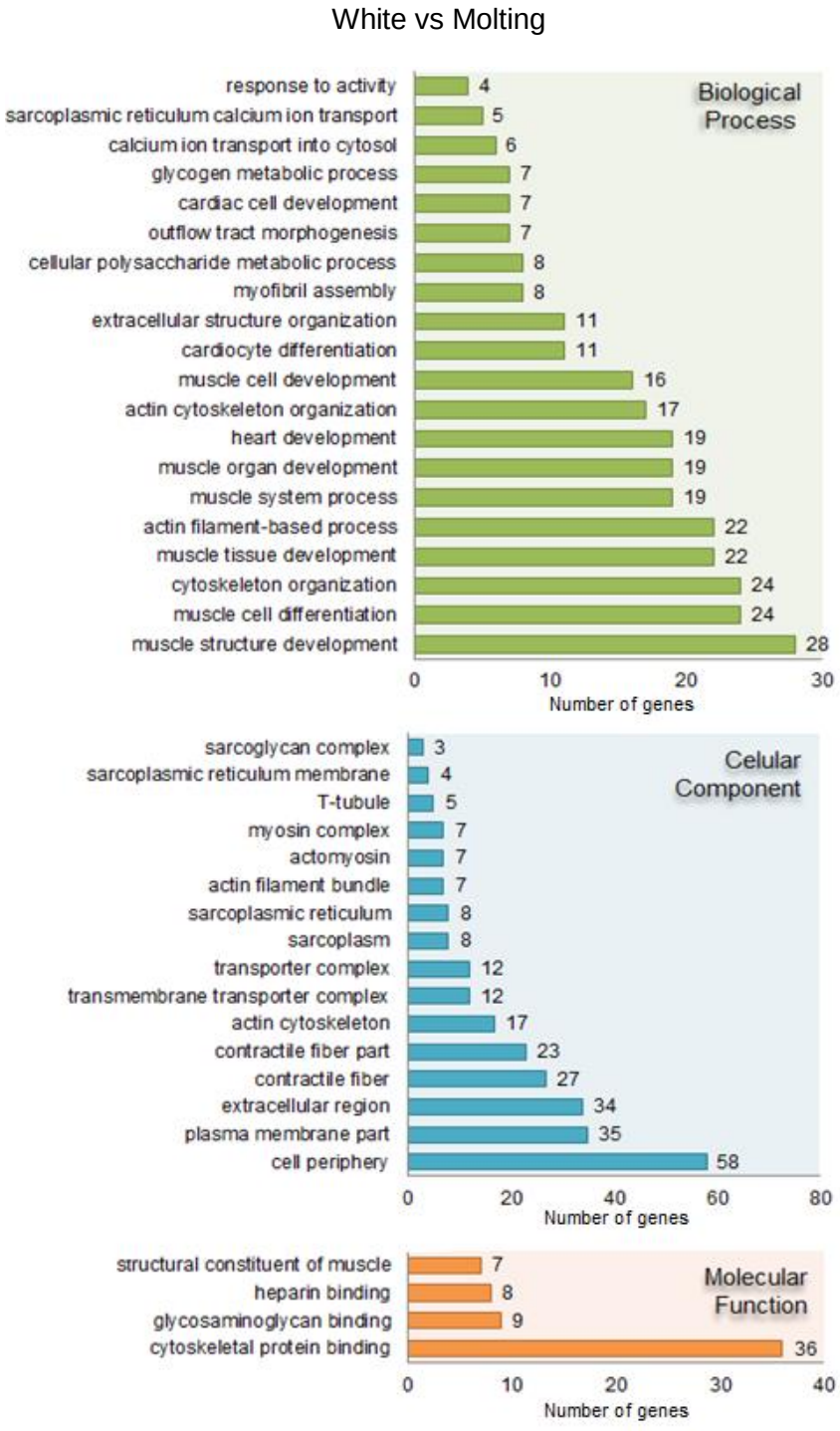

Figure 3.12 – Gene Ontology enriched terms (FDR<0.05) of all GO categories for genes being differentially expressed between “molting” and “white” stage. Numbers next to the bars in the graphs represent the number of genes (in this case, ENSEMBL gene codes) that are associated with the respective GO term.

The main aim of this analysis was to find which categories could be related with the coat color phenotype, and those related with skin development and morphogenesis are obvious candidates. Such categories are only significantly enriched in DE genes between “brown” and “white” stage (Figure 3.14; Appendix 1 – Table S3), with “hair follicle development” and “epidermis morphogenesis” being significantly enriched with 8 and 9 genes respectively (pointed with red arrows in Figure 3.14). However, this analysis allowed identifying additional categories that might be involved in the process.

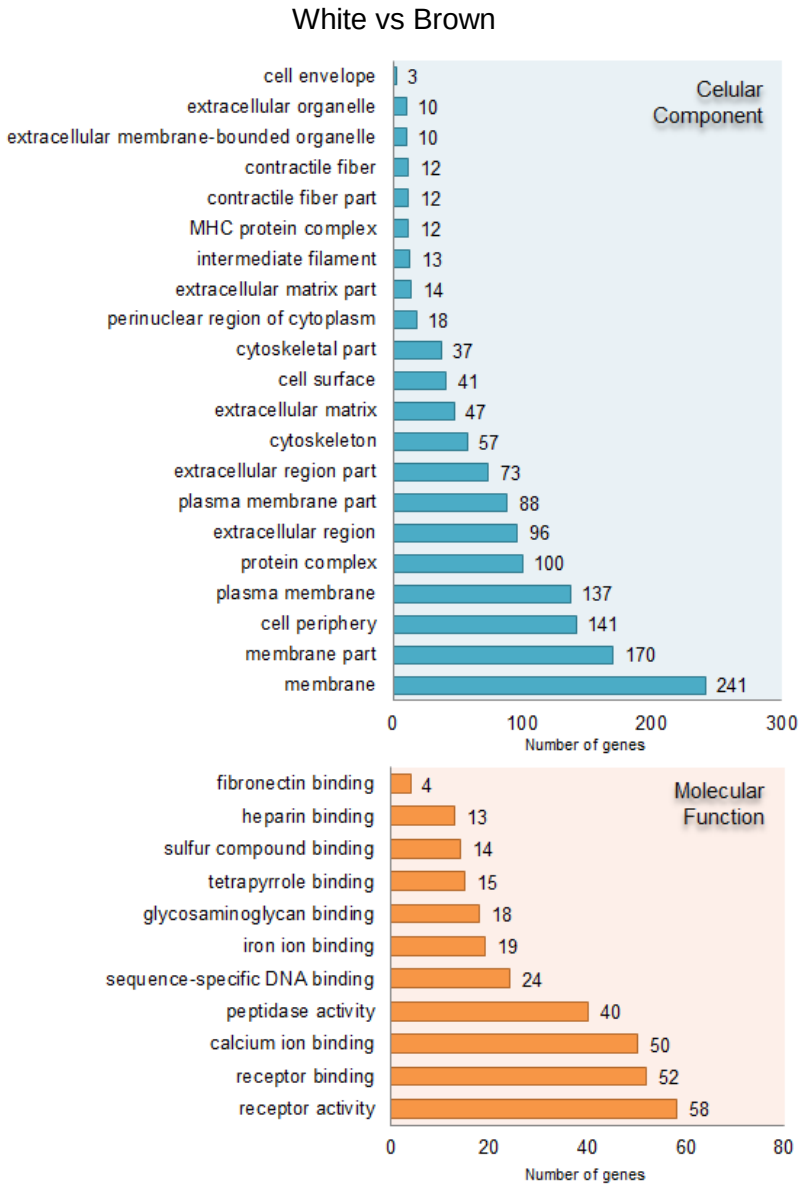

Figure 3.13 - Gene Ontology Celular Component and Molecular Function enriched terms (FDR<0.05) and categories for genes being differentially expressed between “white” and “brown” stage. Numbers next to the bars in the graphs represent the number of genes (in this case, ENSEMBL gene codes) that are associated with the respective GO term.

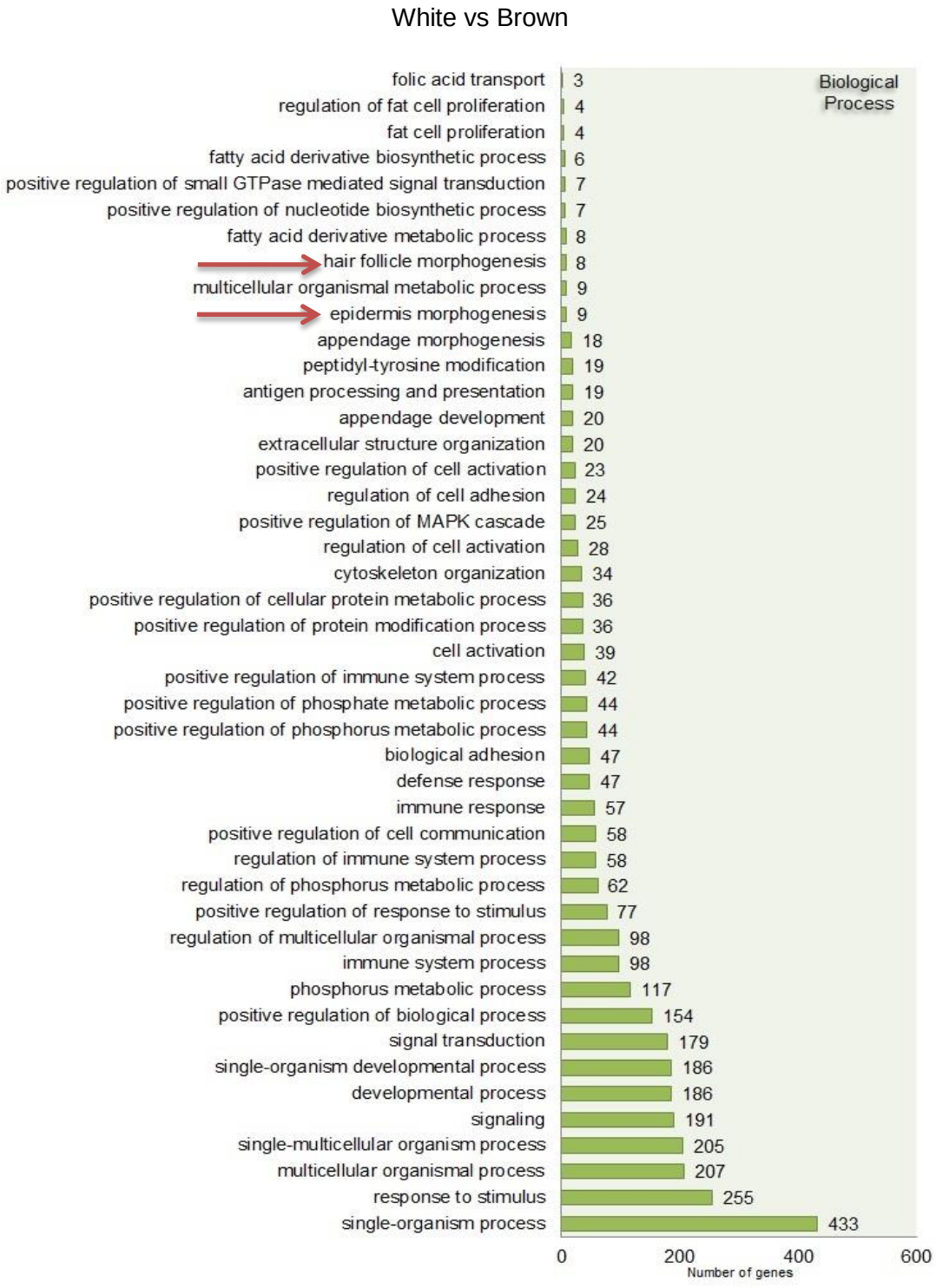

Figure 3.14 - Gene Ontology Biological Process enriched terms (FDR<0.05) and categories for genes being differentially expressed between "white" and "brown" stage. Numbers next to the bars in the graphs represent the number of genes (in this case, ENSEMBL gene codes) that are associated with the respective GO term. Red arrows point out the two GO terms related with skin development.

A closer look at the list of genes associated with the terms reveals that genes associated with “hair follicle development” and “epidermis morphogenesis” are actually the same except one (Table 3.4), which is logical since “hair follicle morphogenesis” is a child term of “epidermis morphogenesis” in the GO annotation. In Table 3.4, the component codes annotated to the ENSEMBL gene codes and the gene names associated with them are listed. One of the components is annotated to two ENSEMBL gene codes which, as explained before, is due to the fact that the isoforms of this component were annotated to different genes. The expression patterns of these components were drawn from the clusters they fell in in the partition cluster analysis referred before. They either fell in cluster 2 (Figure 3.15) or cluster 3 (Figure 3.16) meaning that their expression generally either increased or decreased along the molt, respectively.

Table 3.4- List of ENSEMBL gene codes in each of the GO terms in bold. For each ENSEMBL code the respective Trinity component code and associated gene name is listed. All of the components are differentially expressed only in “white” vs “brown” pairwise comparison

Hair follicle morphogenesis	**Epidermis morphogenesis**		
ENSEMBL gene code		Component code	Associated gene name
ENSOCUG00000008499	ENSOCUG00000008499	comp135860_c0	*FGF10*
ENSOCUG00000008992	ENSOCUG00000008992	comp152336_c2	*KRT25*
ENSOCUG00000009003	ENSOCUG00000009003	comp152485_c0	*KRT27*
ENSOCUG00000013731	ENSOCUG00000013731	comp152403_c1	*RUNX1*
ENSOCUG00000016465	ENSOCUG00000016465	comp152107_c1	*TP63*
ENSOCUG00000017751	ENSOCUG00000017751	comp152403_c1	*RUNX3*
ENSOCUG00000025513	ENSOCUG00000025513	comp145734_c0	*IGFBP5*
ENSOCUG00000029234	ENSOCUG00000029234	comp149548_c2	*KRT71*
	ENSOCUG00000021812	comp152197_c0	*PLOD3*

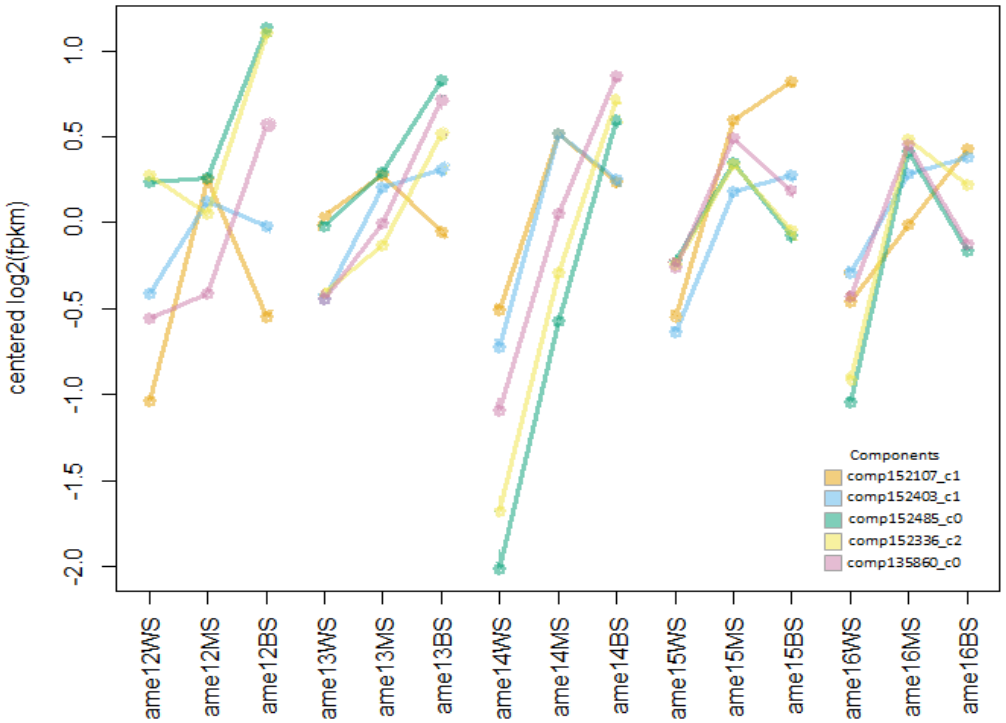

Figure 3.15 – Expression levels for each of the Trinity components showing an overall increase in expression and associated with the GO terms "hair follicle development" and "epidermis morphogenesis", in each sample. In the xx axis, in the individual codes of each sample the "molting" stage is symbolized as WS ("white"), MS ("molting"), BS ("brown").

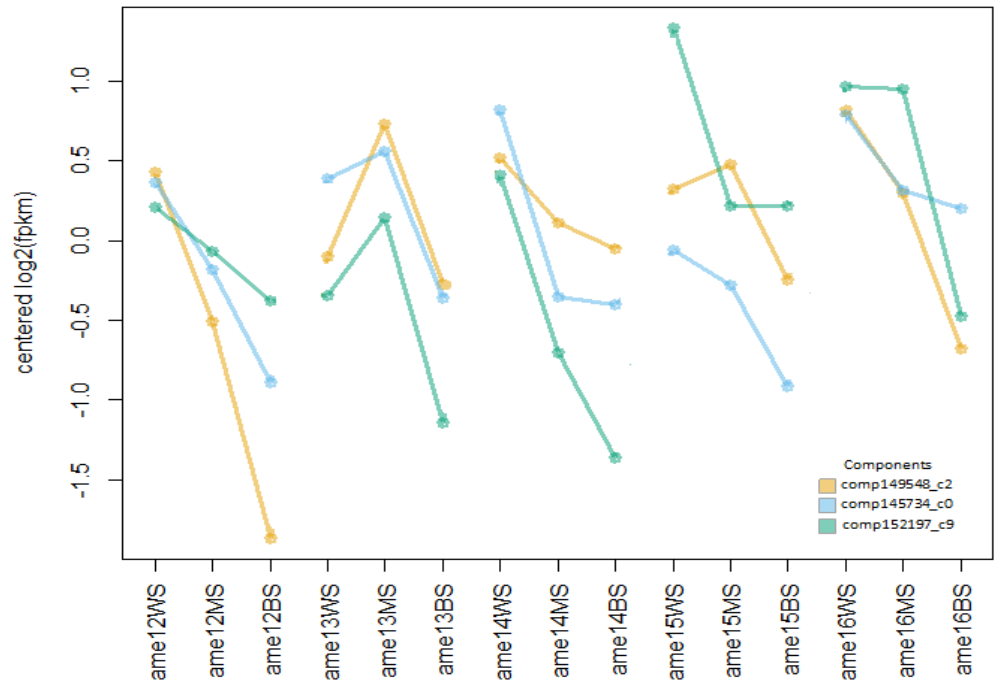

Figure 3.16 - Expression levels for each of the Trinity components showing an overall decrease in expression and associated with the GO terms "hair follicle development" and "epidermis morphogenesis", in each sample. In the xx axis, in the individual codes of each sample the "molting" stage is symbolized as WS ("white"), MS ("molting"), BS ("brown").

3.6. Candidate gene perspective

The candidate gene perspective allowed finding six cases related with the functions analyzed. Some of these genes were associated with the GO term "response to stimulus" such as the genes coding for the agouti signaling protein (ASIP), melanocortin-4 receptor (MC4R), nuclear receptor Rev-erbα (NR1D1), retinoic acid receptor–related orphan receptor (RORα) (Table 3.5; Appendix 1 – Table S3). The calcium-activated potassium channel subunit α-1 (RBSLO1/KCNMA1) has been implicated in the circadian rhythm regulation (Meredith *et al.*, 2006). This gene was associated with Biological Process GO terms such as "muscle system development" and Cellular Component GO term "transmembrane transporter complex" (Appendix 1 – Table S5). Myosin type VIIa (MYOVIIA) encoding gene was the only one related with melanosome transport encountered in the dataset (Table 3.5). This gene was associated with GO terms such as Biological Process term "developmental process" and Cellular Component term "cytoskeleton" (Appendix 1 – Table S3). The expression pattern of these genes decreased towards the end of the molt and therefore they fall in the "cluster 3" of the partitioning clustering analysis (Figure 3.17).

Table 3.5 - List of components which are associated with "response to stimulus" GO term category (NRD1, RBSLO1, RORα, ASIP, MC4R), transporters of keratinocytes (MYOVIIA) and melanogenesis (ASIP) and their respective component code, ENSEMBL gene code, associated gene name, pairwise comparison where they are differentially expressed, and the genes pairwise comparison associated with significantly enriched GO terms. BSWS – Brown vs White; MSWS – Molting vs White

Component code	ENSEMBL gene code	Associated gene name	Pairwise comparison	GO terms of DE genes in the pairwise comparison
comp140510_c3	ENSOCUG00000009732	*NR1D1*	BSWS	BSWS
comp152200_c0	ENSOCUG00000009543	*RBSLO1/KCNMA1*	MSWS	MSWS
comp140802_c4	ENSOCUG00000012900	*RORα*	BSWS	BSWS
comp146368_c0	ENSOCUG00000025457	*MC4R*	BSWS	BSWS
comp152787_c1	ENSOCUG00000023355	*MYOVIIA*	BSWS	BSWS
comp110590_c0	ENSOCUG00000011138	*ASIP*	BSWS;MSWS	BSWS

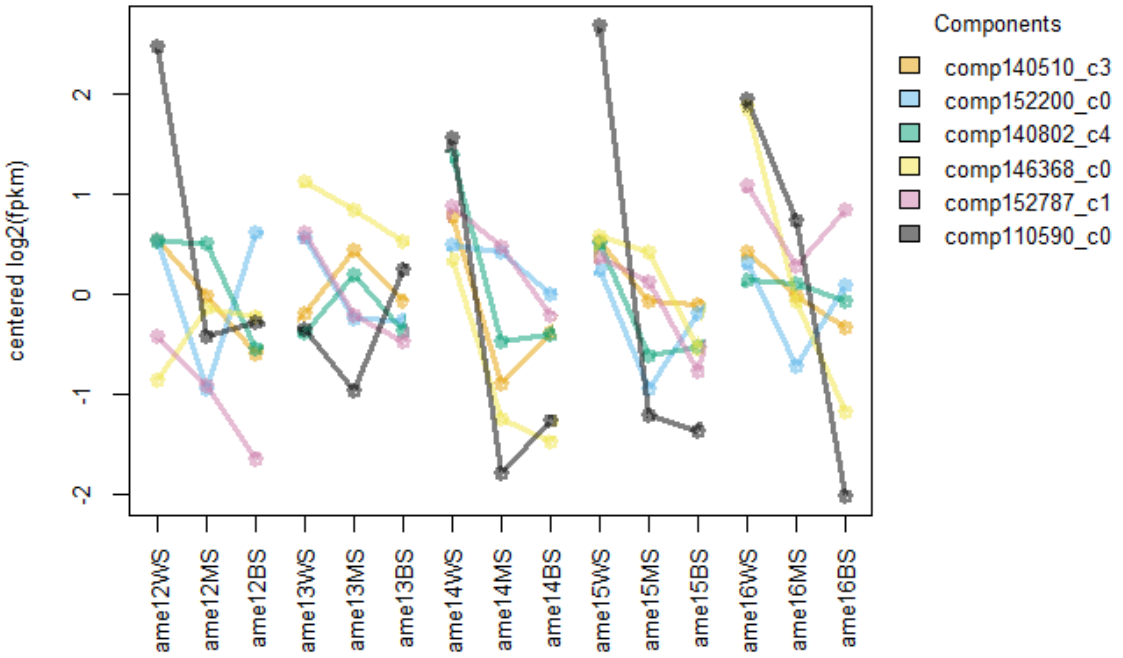

Figure 3.17 - Expression levels for each of the Trinity components associated with the genes in Table 3.5. In the xx axis, in the individual codes of each sample the “molting” stage is symbolized as WS (“white”), MS (“molting”), BS (“brown”).

4. Discussion

Seasonal coat color change translates an important adaptation in arctic and boreal species. It allows these species to match the seasonal change in background color due to snow fall and thus to conceal from other organisms, namely predators, *i.e.* crypsis. Although there is some evidence for the hormonal control and, in the light of some studies, it is possible to rule out or to advance possible candidate genes for the seasonal coat color change, the genetic mechanism responsible for the change is unknown. Changes in gene expression are known to be one of the mechanisms that allow species to adapt to different local environments (Larsen *et al.*, 2007). Furthermore, gene expression studies have also been applied to study time-course gene expression and understand which genes might be involved in specific phenotypes (Schlake *et al.*, 2004; Lin *et al.*, 2004; Stoughton, 2005; Margolin *et al.*, 2014). In this study, RNA-seq, a high throughput sequence technology, was successfully applied to obtain an initial snapshot of the transcriptional changes occurring during the spring molt in the snowshoe hare. The full transcriptome of the skin was assembled and used as reference for mapping and estimation of relative abundances. Furthermore, this method allowed the identification of differentially expressed genes at different molt stages, and the expression patterns of these genes might elucidate the transcriptional changes occurring in the molting skin. Finally, functional categories associated with the DE genes have validated the potential of the methodology to capture the molecular machinery involved in skin-related processes such as hair growth cycle and epidermis development and opened the possibility to draw a list of candidates related with seasonal melanogenesis.

4.1. Bioinformatics pipeline

One of the objectives of this work was to set up and apply a bioinformatics analytical pipeline to uncover differentially expressed genes in different skin types during the coat color molt. It is thus important to discuss the robustness of the applied methodology and possible improvements that can be introduced in the future, particularly taking into account the difficulties of *de novo* transcriptome assembly.

Despite allowing the discovery of new transcripts and opening the door for high throughput analyses in non-model organisms without a reference genome (Martin and Wang, 2011), which dramatically increases the number of cases where expression studies are possible, *de novo* assemblies have some associated difficulties. Among them is the generation of dubious contigs due to lack of coverage in some regions or due to the existence of errors or highly similar regions such as isoforms or paralogous (Cahais *et al.*, 2012; Singhal, 2013). Quality trimming prior to the assembly is one of the strategies used to remove errors related with sequencing or library generation and has been proven to increase the quality of the dataset and assembly (Del Fabbro *et al.*, 2013; Singhal, 2013). In this work, this step was not overlooked and low quality reads and adaptors were removed resulting in the reduction of ~15% in the number of initial read pairs. Also, coverage/sequencing depth is an important aspect influencing the quality of *de novo* assemblies. Here, however, deep sequencing was performed (data from 2+2 lanes of data from Illumina HiSeq2000 and 2500 was collected), exceeding what is deemed enough for a reliable transcriptome assembly (Vijay *et al.*, 2013).

The raw assembly (previous to the filtering step implemented) yielded a large number of contigs (321441). When compared to other transcriptome assemblies performed in *Lepus* with different sequencing and assembly strategies (38540 contigs in Cahais *et al.*, 2012; below 85000 contigs in Gayral *et al.*, 2013) or with other assemblies performed with Trinity with a similar sequencing strategy (same sequencer and number of lanes) in other organisms (from ~84000 to ~113000 contigs in Singhal, 2013), this number is higher. Filtering out non-annotated contigs lowered the number of contigs (85967) but still resulted in 4 times the number of protein coding genes annotated in mouse, human and rabbit, in the ENSEMBL database (22740, 22836 and 19293, respectively). Thus the numbers of contigs in the transcriptome seem higher than what expected, despite providing high completeness and contiguity (N50). The filter for coverage finally reduced the number of components to 14568, suggesting that these filtering steps allowed recovering a realistic number of genes when compared to the reference. However, some Trinity components (that may represent genes) had multiple annotations and improvements can be made in the pipeline followed in this work. Because RSEM sums isoform counts to obtain the respective components relative abundances, this fact has an implication in the interpretation of gene expression. If one component corresponds in fact to several genes this means that Trinity was not completely successful in classifying the isoforms of multiple genes into different components. In the future, the codification of these components may

be revised based on the annotation to ensure that each component corresponds only to one gene annotated by blastx hit.

The above properties of the generated assembly show that *de novo* assemblies should be carefully analyzed and undergo several filtering steps (Singhal, 2013). Additional filters may be introduced in the pipeline. For example, completeness results may be used to filter contigs that did not cover the reference more than a specific length. Also a more tight Blastx analysis (e.g. lower e-value or imposing a minimum length of blastx hit) may be used to reduce the probability of annotation of erroneous contigs. Alternatively, the rabbit transcriptome could be used as reference both for mapping and annotation in a parallel pipeline of analysis. Even though this would introduce other sorts of biases in the data that we wished to avoid here, such as lower efficiency and quality of mapping or bias towards mapping in conserved genes, a comparison of the strategies could provide valuable insights on the robustness of the analyses.

4.2. Gene expression in the skin along the spring molt

4.2.1. Individual variation in gene expression

The multidimensional scaling plot revealed a consistent log fold change among expression estimates resulting from different sequencing lanes (Figure 3.2a). However, the relative sample location in the second plot of Figure 3.2 reveals both a transcriptional similarity between molting stages and the existence of individual variability, which is likely due to normal biological variation. Indeed, microarray expression profiling in fish have shown that gene expression patterns between individuals can be as high as differences between tissues, finding that 48% of the total genes assayed were differentially expressed among individuals (Whitehead and Crawford, 2005). The same study also reported that some genes are not consistently expressed between tissues within the same species. The time of day or age can also influence gene expression in humans (Whitney *et al.*, 2003), and sex has been implicated to shape gene expression in fruit flies (Jin *et al.*, 2001). Since RNA-seq has the potential to capture the complete population of mRNA molecules in the skin, it is likely that some inconsistent patterns between individuals may be found.

Variation between biological replicates was 22% but a great proportion of the genes had higher BCV (tagwise) estimates (Figure 3.3). Therefore, variation between individuals in gene expression is expected and the observed differences likely result from successfully capturing such variation.

4.2.2. Differences in gene expression among molting stages

Despite the apparent similarities in log fold changes between molting stages, this work allowed identifying differentially expressed genes between them. Reflecting the closer proximity between “brown” and “molting” stages in the MDS plot (Figure 3.2b) fewer genes were differentially expressed between these two stages. This might be explained by a similarity of the expression patterns between “brown” and “molting” samples, as suggested by the hierarchical clustering. Three out of five “molt” samples cluster closer to “brown” samples. Nevertheless, a similar number of “molt” samples, two, cluster with “white” samples, which could be explained by some inconsistencies in the exact biopsy location or variation in the phase of the molting cycle among specimens.

On the other hand, the pairwise comparisons between “white” vs “brown” and “white” vs “molting” produced the highest and the second highest number of differentially expressed genes, respectively. This is in agreement with the hierarchical clustering analysis since “white” samples clustered together with two exceptions. This implies a higher difference between the expression levels of “white” and “brown” and “white” and “molting”, with “white” being the stage with the highest levels of expression. Reflecting this, expression tended to progress like a wave, with most genes being highly expressed in white and lowering their expression towards the end of the molt. The opposite (increased expression) happened to a relative smaller proportion of genes (~29%). The molting skin thus undergoes major transcriptional changes, with different groups of genes being highly expressed at different stages and likely performing different functions, in conformity with the phenotypic changes observed. This is consistent with the switching and silencing of genes that are related with new hair growth (Stenn and Paus, 2001; Schlake *et al.*, 2004; Lin *et al.*, 2004). In fact, it is encouraging to have captured the genetic machinery involved in hair growth in this analysis. The hair growth cycle is divided in three stages: active growth (anagen), regression (catagen) and quiescence (telogen) (Stenn and Paus, 2001).

Anagen is characterized by the formation of a new hair follicle by active proliferation of cells of the secondary hair germ. These eventually differentiate the different layers of the follicle that produce the hair shaft (Stenn and Paus, 2001). Melanogenesis is restricted to anagen and once anagen ends both melanogenesis and hair growth stop (Slominski and Paus, 1993; Stenn and Paus, 2001). With the end of anagen, growth stops and the hair follicle regresses. Catagen is thus characterized by cessation of proliferation, cell apoptosis and cell structural changes that lead to a shrinkage of the follicle (Stenn and Paus, 2001). Telogen follows catagen and at this stage the cells at the follicle are quiescent and no synthesis of DNA or RNA occurs (Stenn and Paus, 2001). The hair shaft remains attached to the follicle and eventually sheds when anagen restarts or when it naturally falls during the animals daily activities (Paus and Cotsarelis, 1999). Since no brown hair is observed in the “white” stage this phase likely corresponds with late stages of telogen of the previous hair cycle or early to mid anagen. The emergence of brown hair that occurs in “molting” likely corresponds to anagen and possible the start of catagen and “brown” late catagen and telogen. Studies that followed gene expression during hair growth cycle reported that late telogen and anagen are the most transcriptionally active stages, with genes involved in hair growth, cell proliferation and encoding structural proteins being expressed (Schlake *et al.*, 2004; Lin *et al.*, 2004). This would explain the high levels of expression in the “white” stage. Also, both studies report another peak in expression in late catagen and telogen that could also be the peak observed in the “brown” stage. In the hair growth cycle, this increase of expression in the end of the cycle seems to be related with the expression of genes involved in apoptosis, immune response, and categories related with the stabilization of the tissue in the regression phase of hair growth (Schlake *et al.*, 2004; Lin *et al.*, 2004).

Overall, differentially expressed genes followed three expression patterns during the spring molt cycle, as observed in Figure 3.9. Genes’ expression either increases throughout the molt, decreases, or follows an intermediate pattern where the “molting” stage is the lower point of expression. According to what was already observed in the hierarchical clustering result, the greatest proportion of the genes decreased their expression towards the end of the molt (Figure 3.9, Cluster 3) and less genes followed the opposite pattern (Figure 3.9, Cluster 2). It is possible that although k was estimated to be 3, more patterns could have been drawn if larger numbers of k were used, similar to the results of Lin *et al*. (2004). Nevertheless, these patterns seem related with the expression of genes in the different stages of the hair growth cycle, in parallel to what was suggested

before. For instance, Lin *et al*., (2004) also suggested the existence of three main categories that are related with the different phases of the hair growth – anagen, catogen and telogen.

4.2.3. Functional genetic context of the spring molt

Gene Ontology annotation was used to find which functions were significantly enriched in the set of differentially expressed genes. The enriched functions were related with immunity, muscle tissue and structure development, metabolism ("glycogen metabolic process"), appendage development, skin morphogenesis and terms related with cellular component organization such as cytoskeleton organization. Many of the terms significantly enriched in the skin can thus be related with the growth of new hair and the changes that are occurring not only in the follicle but also in the skin to promote that growth (Lin *et al.*, 2004). Because there is a massive regeneration of hair follicles during the molt, enrichment of terms related with cytoskeleton organization or cell activation result from the fact that cells are proliferating at the early stage of hair growth (Lin *et al.*, 2004). Moreover, Biological Function GO terms related with immunity are enriched considering DE genes between the "brown" and "molting" stages (Figure 3.11) and "brown" and "white" stages (Figure 3.14), but not between "white" and "molting" stages (Figure 3.12). The involvement of immunity-related genes in the hair growth cycle (Stenn and Paus, 2001) and also in pigmentation and melanocyte function (Slominski *et al.*, 2004) has been reported. Furthermore, during hair growth apoptotic events occur that may trigger inflammatory reactions from the immune system (Stenn and Paus, 2001; Schlake *et al.*, 2004). Genes involved in immune response being expressed in the skin can be related with the death of the animal (hares were euthanized to collect other organ samples) that produced an inflammatory response or with an injury that the animal had at the time of death (Pritchard *et al.*, 2001). On the other hand, genes such as interleukin 1 (IL-1) and 6 (IL-6) have been found to inhibit melanogenesis in humans (Swope *et al.*, 1991). Furthermore, terms related with fat cell proliferation and regulation might be related with the formation of adipocytes in the skin. Recent evidences point to the fact that adipocytes in the skin might play a role of the hair growth cycle (Festa *et al.*, 2011), what would make sense in the skin's transcriptional window being analyzed here.

On another note, DE genes between "molting" and "white" stage were significantly related with muscle, heart and cell developmental functions (Figure 3.12). Interestingly, the cellular component term "myosin complex" is significantly enriched. In the context of the skin and melanogenesis, myosines, which are mechanoenzymes that generate movement in the cell and are important for the transport via the actin cytoskeleton of cellular components (reviewed in Kneussel and Wagner, 2013), and in particular myosin Va (MYOVA) and myosin VIIa (MYOVIIA) are important to mediate the transport of melanosomes into keratinocytes and into the retinal pigmented epithelium, respectively (Liu *et al.*, 1998; Wu *et al.*, 1998).

The GO enrichment analysis suggested significant enrichment of terms related with hair follicle morphogenesis and epidermis morphogenesis and so a closer look into the genes related with these terms was taken (Table 3.4; Figure 3.15-3.16), which provided further insight into the hair growth cycle machinery captured in this work. For instance, fibroblast growth factor 10 (FGF10) belongs to the fibroblast growth factor gene family having high homology and functional similarities with FGF7 and is known to be expressed in human and mouse skin (Beer *et al.*, 1997) and to stimulate keratinocyte differentiation in *in vitro* cultured human cells (Marchese *et al.*, 2001). Both FGF7 and FGF10 are implicated in the initial steps of the hair growth cycle (Greco *et al.*, 2009). On the other hand, there is evidence that FGF10 and 7 control the production of Insuline-Like Growth Factor binding protein 5 (IGFBP5) via another receptor and this interferes with the hair shaft medulla formation (Schlake, 2005a). Tumor Protein p63 (TP63) is a transcriptional factor that has been implicated in the regulation of other genes responsible for the normal growth and stratification of the skin (McDade *et al.*, 2012). Furthermore, both Runt-related transcription factor 1 (RUNX1) and Runt-related transcription factor 3 (RUNX3) are transcriptional regulators expressed in the dermal compartment of hair follicles and are implicated in the under-fur hair shape, since knockout mice presents less bents than the wild type zigzag typical shape (Raveh *et al.*, 2005, 2006); RUNX3 was also found to be expressed in melanocytes, although its function there is not known (Raveh *et al.*, 2005). Interestingly, IGFBP5 has been hypothesized to be an effector of RUNX1 and RUNX3 because of its role in the formation of the bends characterizing zigzag hair types (Schlake, 2005b). Keratinocytes produce keratin which is a protein encoded by genes such as *KRT25, KRT27* and *KRT71*. These genes are expressed in the inner root sheath of the hair follicle and *KRT25 and 27* are also expressed in the hair medulla (Langbein *et al.*, 2006). Interestingly, Lin *et al.* (2014) showed that the expression of keratin-associated genes

peaked in the late anagen of the growth cycle and, similarly, here the expression of *KRT27* and *KRT25* peaked towards the end of the molt. Finally, Procollagen-lysine, 2-oxoglutarate 5-dioxygenase 3 (*PLOD3*) gene, the only gene only associated with epidermis morphogenesis, is a lysyl hydroxylase and is important for the proper formation of collagen, a component of the epidermis (Uzawa *et al.*, 1999). The expression of these genes is consistent with the fact that a major change takes place in the skin since new hair is growing during the molt change. Although hypothesizing the role of any of these genes in the coat color change would be premature, given the results obtained and given their functions, it is remarkable to have captured the expression of these genes and ultimately the machinery involved in the hair cycle. For example, FGF10 presumably promotes the expression of another receptor that negatively regulates the production of IGFBP5 and, in fact, while an increase in the expression of FGF10 is observed along the molt, the reverse is observed for IGFBP5 (Schlake, 2005a). In parallel, it was observed an increase in expression of *FGF10* (Figure 3.15) while a decrease for *IFGFP5* was captured (Figure 3.16).

4.2.4. Genes involved in melanogenesis and circadian rhythm

This work was able to capture the machinery involved in the hair growth cycle, and a closer look at DE genes related with melanogenesis, melanosome transport and circadian rhythms, which could be involved in the seasonal coat color change. Synthesis of melanin is restricted to anagen (Slominski and Paus, 1993), and therefore it is not surprising to find that genes known to be related with melanogenesis are mostly differentially expressed between “white” and “brown” or “molting” and “white”, and have a decreasing pattern of gene expression towards the end of the molt (Table 3.5; Figure 3.17). The agouti signaling peptide (*ASIP*) gene is one of the components of the melanin synthesis pathway and was found here to be expressed at high levels in the beginning of the molting cycle. This peptide is responsible for the production of agouti hair, characterized by brown hair with a yellow band. This is achieved by the antagonistic binding to melanocortin-1 receptor (MC1R) what temporarily impedes the binding of α-MSH (Vrieling *et al.*, 1994). In the rabbit, variants of *ASIP* have been associated with different color phenotypes and the wild variant of *ASIP* (A^{w}) in rabbit is responsible for the production of hair similar to the wild type A allele know as white-bellied agouti in mice (Fontanesi *et al.*, 2010). Hares’ spring

coat is brown except on the belly and therefore ASIP expression is likely regulating the production of brown hair. MC4R is also a target of ASIP and is associated with regulation of body size both in humans (Yeo *et al.*, 1998), mice (Huszar *et al.*, 1997) and rabbit (Fontanesi *et al.*, 2013). Although it has mainly a role in regulating body size, it is also expressed in the human melanocytes and keratinocytes and some evidence exists that it might be also a regulator of melanogenesis (Spencer and Schallreuter, 2009). Furthermore, Myosin VIIa (MYOVIIA), as referred previously, is a transporter of melanosomes and mice having a mutant version of this gene have abnormal distribution of melanosomes in the retina pigment epithelium (Liu *et al.*, 1998). Myosin VIIa mutant mice display mainly hearing and equilibrium defects (Hasson *et al.*, 1997).

The nuclear receptor Rev-erbα (NR1D1), the retinoic acid receptor–related orphan receptor (RORα) and the calcium-activated potassium channel subunit α-1 (RBSLO1/KCNMA1) are gene products associated with regulation of circadian rhythms in mammals (Akashi and Takumi, 2005; Meredith *et al.*, 2006; Yin *et al.*, 2010). The circadian clock in the mammal system results from positive and negative feedback loops (Reppert and Weaver, 2002). Therefore, the CLOCK-BMAL1 heterodimer promotes the expression of targets, such as NR1D1 which in turn, together with RORα, promotes a decrease in expression of *BMAL1* (Preitner *et al.*, 2002; Akashi and Takumi, 2005). The circadian clock genes have been shown to be expressed in the skin, playing a role in the regulation of hair cycling (Lin *et al.*, 2009). The expression of genes such as *NR1D1* was shown to significantly peak in telogen (also non significantly in late anagen) (Lin *et al.*, 2009) which is in concordance with what was found in this study, if the “white” stage is assumed to be the proxy of late telogen and anagen. RORα isoforms, on the other hand, were found to be expressed in human skin cells such as dermal fibroblasts, epidermal melanocytes and immortalized epidermal keratinocytes (Slominski *et al.*, 2005). Another study found that RORα is mostly expressed from mid-anagen to telogen during the hair-growth cycle (Kobayashi *et al.*, 2005), which does not closely agree with the expression profile found in this study, even though this experiment was not designed to capture in detail the hair growth cycle expression (Figure 3.17). Furthermore, it has been hypothesized that RORα is either a direct or indirect receptor of melatonin (Slominski *et al*., 2012 and references therein). Kobayashi *et al.* (2005) showed that the expression of RORα increased during the same stages in which melatonin was detected in the hair follicle. Therefore, this could be an integrator of the melatonin signal in the melanocytes of the skin as also hypothesized by others (Kobayashi *et al.*, 2005; Slominski *et al.*, 2012). Finally, KCNMA1

knockout (Kcnma1–/–) mice show a decrease amplitude of the circadian-associated behaviors and the authors proposed that these K^+ channels might be mediators of the circadian clock output (Meredith *et al.*, 2006). The gene has been found to be expressed in human melanoma (Mazar *et al.*, 2010). Interestingly, KCNMA1 is the target of the miRNA miR-221, and abnormal low expression of miR-221 in the melanoma human cells studied by Mazar *et al*. (2010) induced an increase in KCNMA1 expression. The gene encoding this miRNA is, on the other hand, regulated by microphthalmia associated transcription factor (MITF), which is known to be involved in the melanogenic pathway (Gaggioli *et al.*, 2003).

4.3. Future prospects

Since the molt involves the replacement of the hairs in the body of the animal (Severaid, 1945), having captured functional categories and some genes involved in this process can be regarded as a sign of the success of this assay. Furthermore, the specific genes related with melanogenesis, circadian rhythm and melanosome transport analyzed here suggest that this approach has the potential to be used to study seasonal coat color change even further. Having captured these examples, it is possible to imagine the potential of a sampling strategy particularly designed to follow the hair growth cycle during the molt to capture hair growth cycle and molt's transcriptional changes. A more detailed sampling similar to the ones made in gene expression studies that follow the hair growth expression in day intervals could grant a better capture of the genes involved in the pigmentation pathway (see as examples Lin *et al*., 2004, 2009; Schlake *et al*., 2004). Having captured genes in the skin such as Myosin VIIA that have a functional role in pigmentation in other organs but still are expressed even if at low levels suggests that other relevant genes could have been captured either with a more detailed sampling or by improving the annotation in this pipeline. Because the number of putative genes analyzed in the end (14568) was inferior to the total number of protein coding genes in the reference (23669) it is possible that some important genes for the phenotype under study, such as *MC1R*, *TYR, TYRP1* or *TYRP2*, were not retained in the analysis either because they were filtered out during the pipeline or are not annotated in the reference. Therefore, the analysis can likely profit from, for example, an annotation with human or mouse references.

Also, it remains to zoom in further into the functional categories and to paint a detailed picture of the genes and gene pathways from the transcriptional snapshot taken here. Other interesting cases to further explore would be genes expressed in specific stages. For example, the 6 genes differentially expressed in the "molting" stage may be involved in the trigger leading to the seasonal molt. Moreover, a comparison with the autumn molt transcriptome would grant a deeper insight into the seasonal coat color change since the gene expression of the machinery of hair growth cycle and skin related functions would be controlled for. Most likely, genes that are involved in the production of the brown coat will have a different pattern of expression when a white coat is being produced. Also, studying polymorphic populations with individuals that either change or not to a white winter phenotype would allow to compare both autumn and spring transcriptome while controlling for the underlying background of expression changes, further reducing the list of candidate genes for the seasonal coat color change. Ongoing research in this system aims at taking advantage of the existence of these populations to identify genes associated with the variable phenotype and ultimately test the action of selection in specific regions of the snowshoe hare genome. Finally, comparative transcriptomic analysis between snowshoe hares and other hare species that also seasonally change their color can provide an evolutionary perspective on the phenotype. In this perspective, future work will focus on the mountain hare (*L. timidus*) using the promising strategy described here. This will allow understanding the (parallel) evolution of seasonal coat color change.

5. Conclusion

In this study an insight into the transcriptional changes occurring during the spring molt of the snowshoe hare (*L. americanus*) was provided. The results obtained demonstrate the power of the new sequencing technologies to capture gene expression changes along cyclic processes such as seasonal coat color change. These can be applied to the study of non-model species and here allowed building the skin transcriptome of the snowshoe hare and to detect the major transcriptional waves occurring during the process. It was also possible to identify the major functional categories of the genes involved in the molt, such as immune response, muscle tissue and structure development, metabolism, appendage development, skin morphogenesis and terms related with cellular component organization such as cytoskeleton organization, and to understand how the expression of genes related with the pigmentation cycle vary along the molt, such as *ASIP*, *MC4R* and *MYOVIIA* and genes involved in producing a circadian rhythm (*NR1D1, RORα, RBSLO1*).

Snowshoe hares, as other artic species that mimic the background color of their environments by seasonally changing the color of their coats and feathers, will likely be strongly affected by global warming, which is altering the duration of snow cover. In this context, studies as the one developed here that attempt to understand the molecular basis of adaptive phenotypes are important to provide the tools to determine the potential for adaptation or plasticity. They are a vital step onto the understanding of the ability of species to accompany major effects of global warming on boreal species, and ultimately to adapt.

6. References

Adamík P, Král M (2008). Climate- and resource-driven long-term changes in dormice populations negatively affect hole-nesting songbirds. *Journal of Zoology* **275**: 209–215.

Akashi M, Takumi T (2005). The orphan nuclear receptor RORα regulates circadian transcription of the mammalian core-clock Bmal1. *Nature Structural & Molecular Biology* **12**: 441–448.

Alves PC, Hackländer K (2008). Lagomorph Species: Geographical Distribution and Conservation Status. In: Alves PC, Ferrand N, Hackländer K (eds) *Lagomorph Biology: Evolution, Ecology and Conservation*, Springer: Berlin, Germany, pp 395–405.

Anderson T, Candille S, Musiani M (2009). Molecular and evolutionary history of melanism in North American gray wolves. *Science* **323**: 1339–1343.

Andrews S (2014). FastQC A Quality Control tool for High Throughput Sequence Data. *http://www.bioinformatics.babraham.ac.uk/projects/fastqc/*.

Angelini C, Canditiis D, Feis I (2014). Computational approaches for isoform detection and estimation: good and bad news. *BMC Bioinformatics* **15**: 1–25.

Angerbjörn A, Flux J (1995). Lepus timidus. *Mammalian Species* **495**: 1–11.

Atwell S, Huang YS, Vilhjálmsson BJ, Willems G, Horton M, Li Y, *et al.* (2010). Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. *Nature* **465**: 627–631.

Barrett RDH, Schluter D (2008). Adaptation from standing genetic variation. *Trends in Ecology & Evolution* **23**: 38–44.

Bauer S, Grossmann S, Vingron M, Robinson PN (2008). Ontologizer 2.0 - a multifunctional tool for GO term enrichment analysis and data exploration. *Bioinformatics* **24**: 1650–1651.

Bearhop S, Fiedler W, Furness RW, Votier SC, Waldron S, Newton J, *et al.* (2005). Assortative mating as a mechanism for rapid evolution of a migratory divide. *Science* **310**: 502–504.

Beer HD, Florence C, Dammeier J, McGuire L, Werner S, Duan DR (1997). Mouse fibroblast growth factor 10: cDNA cloning, protein characterization, and regulation of mRNA expression. *Oncogene* **15**: 2211–2218.

Bennett D, Lamoreux M (2003). The color loci of mice–a genetic century. *Pigment Cell Research* **16**: 333–344.

Best T, Henry T (1994). Lepus othus. *Mammalian Species* **458**: 1–5.

Bissonnette T, Bailey E (1944). Experimental modification and control of molts and changes of coat-color in weasels by controlled lighting. *Annals of the New York Academy of Sciences* **XLV**: 221–260.

Bolger AM, Lohse M, Usadel B (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**: 1–7.

Bouck A, Vision T (2007). The molecular ecologist's guide to expressed sequence tags. *Molecular Ecology* **16**: 907–924.

Bradshaw WE, Holzapfel CM (2001). Genetic shift in photoperiodic response correlated with global warming. *Proceedings of the National Academy of Sciences of the United States of America* **98**: 14509–14511.

Cahais V, Gayral P, Tsagkogeorga G, Melo-Ferreira J, Ballenghien M, Weinert L, *et al.* (2012). Reference-free transcriptome assembly in non-model animals from next-generation sequencing data. *Molecular Ecology Resources* **12**: 834–845.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, *et al.* (2009). BLAST+: architecture and applications. *BMC Bioinformatics* **10**: 1–9.

Castoe TA, Koning APJ de, Hall KT, Card DC, Schield DR, Fujita MK, *et al.* (2013). The Burmese python genome reveals the molecular basis for extreme adaptation in snakes. *Proceedings of the National Academy of Sciences of the United States of America* **110**: 20645–20650.

Cheng E, Hodges KE, Melo-Ferreira J, Alves PC, Mills LS (2014). Conservation implications of the evolutionary history and genetic diversity hotspots of the snowshoe hare. *Molecular Ecology* **23**: 2929–2942.

Chiao CC, Hanlon RT (2001). Cuttlefish camouflage: visual perception of size, contrast and number of white squares on artificial checkerboard substrata initiates disruptive coloration. *The Journal of Experimental Biology* **204**: 2119–2125.

Chuong C (1998). Morphogenesis of Epithelial Appendages: Variations on Top of a Common Theme and Implications in Regeneration. In: Chuong C-M (ed) *Molecular Basis of Epitherial Appendage Morphogenesis*, R.G.Landes Company: Austin, Texas, U.S.A, pp 3–10.

Le Corre V, Kremer A (2012). The genetic differentiation at quantitative trait loci under local adaptation. *Molecular Ecology* **21**: 1548–1566.

Cuadrado M (1998). The use of yellow spot colors as a sexual receptivity signal in females of Chamaeleo chamaeleon. *Herpetologica* **54**: 395–402.

D'haeseleer P (2005). How does gene expression clustering work? *Nature Biotechnology* **23**: 1499–1501.

Dalquest W (1942). Geographic variation in northwestern snowshoe hares. *Journal of Mammalogy* **23**: 166–183.

Dalziel AC, Rogers SM, Schulte PM (2009). Linking genotypes to phenotypes and fitness: how mechanistic biology can inform molecular ecology. *Molecular Ecology* **18**: 4997–5017.

Darwin C, Wallace A (1858). On the tendency of species to form varieties; and on the perpetuation of variaties and species by natural means of selection. *Journal of the Proceedings of the Linnean Society of Longon Zoology* **3**: 45–62.

Duan J, Xia C, Zhao G, Jia J, Kong X (2012). Optimizing de novo common wheat transcriptome assembly using short-read RNA-Seq data. *BMC Genomics* **13**: 1–12.

Duncan M, Goldman B (1984). Hormonal regulation of the annual pelage color cycle in the Djungarian hamster, Phodopus sungorus. II. Role of prolactin. *Journal of Experimental Zoology* **230**: 97–103.

Eddy S (2001). Non–coding RNA genes and the modern RNA world. *Nature Reviews Genetics* **2**: 919–929.

Eklund A, Turner L, Chen P (2006). Replacing cRNA targets with cDNA reduces microarray cross-hybridization. *Nature Biotechnology* **24**: 1071–1073.

Excoffier L, Hofer T, Foll M (2009). Detecting loci under selection in a hierarchically structured population. *Heredity* **103**: 285–298.

Del Fabbro C, Scalabrin S, Morgante M, Giorgi FM (2013). An extensive evaluation of read trimming effects on Illumina NGS data analysis. *PloS ONE* **8**: e85024.

Festa E, Fretz J, Berry R, Schmidt B, Rodeheffer M, Horowitz M, *et al.* (2011). Adipocyte lineage cells contribute to the skin stem cell niche to drive hair cycling. *Cell* **146**: 761–771.

Flicek P, Amode MR, Barrell D, Beal K, Billis K, Brent S, *et al.* (2014). Ensembl 2014. *Nucleic Acids Research* **42**: D749–D755.

Flux JEC, Angermann R (1990). The Hares and Jackrabbits. In: Chapman J, Flux JEC (eds) *Rabbits, Hares and Pikas: Status Survey and Conservation Action Plan*, International Union for Conservation of Nature and Natural Resources: Gland, Switzerland, pp 61–94.

Fontanesi L, Battista D, Olio SD, Fornasini D (2013). A missense mutation in the rabbit melanocortin 4 receptor (MC4R) gene is associated with finisching weight in a meat rabbit line. *Animal Biotechnology* **24**: 37–41.

Fontanesi L, Forestier L, Allain D, Scotti E, Beretti F, Deretz-Picoulet S, *et al.* (2010). Characterization of the rabbit agouti signaling protein (ASIP) gene: Transcripts and phylogenetic analyses and identification of the causative mutation of the nonagouti black coat colour. *Genomics* **95**: 166–175.

Forsman A (1997). Thermal capacity of different colour morphs in the pygmy grasshopper Tetrix subulata. *Annales Zoologici Fennici* **34**: 145–149.

Franks SJ, Hoffmann A (2012). Genetics of climate change adaptation. *Annual Review of Genetics* **46**: 185–208.

Franks SJ, Sim S, Weis AE (2007). Rapid evolution of flowering time by an annual plant in response to a climate fluctuation. *Proceedings of the National Academy of Sciences of the United States of America* **104**: 1278–1282.

Gaggioli C, Buscà R, Abbe P, Ortonne J-P, Ballotti R (2003). Microphthalmia-associated transcription factor (MITF) is required but is not sufficient to induce the expression of melanogenic genes. *Pigment Cell Research* **16**: 374–382.

Gayral P, Melo-Ferreira J, Glémin S, Bierne N, Carneiro M, Nabholz B, *et al.* (2013). Reference-free population genomics from next-generation transcriptome data and the vertebrate-invertebrate gap. *PLoS Genetics* **9**: e1003457.

Gayral P, Weinert L, Chiari Y, Tsagkogeorga G, Ballenghien M, Galtier N (2011). Next-generation sequencing of transcriptomes: a guide to RNA isolation in nonmodel animals. *Molecular Ecology Resources* **11**: 650–661.

Goldman BD (1999). The circadian timing system and reproduction in mammals. *Steroids* **64**: 679–685.

Goldman BD (2001). Mammalian Photoperiodic System: Formal Properties and Neuroendocrine Mechanisms of Photoperiodic Time Measurement. *Journal of Biological Rhythms* **16**: 283–301.

Gordon A (2011). Illumina CASAVA-1.8 FASTQ Filter. Available from http://cancan.cshl.edu/labmembers/gordon/fastq_illumina_filter/

Gower BA, Nagy TR, Stetson MH (1994). Effect of photoperiod, testosterone, and estradiol on body mass, bifid claw size, and pelage color in collared lemmings (Dicrostonyx groenlandicus). *General and Comparative Endocrinology* **93**: 459–470.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, *et al.* (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology* **29**: 644–652.

Greco V, Chen T, Rendl M, Schober M, Pasolli HA, Stokes N, *et al.* (2009). A two-step mechanism for stem cell activation during hair regeneration. *Cell Stem Cell* **4**: 155–169.

Griffin PC, Griffin SC, Waroquiers C, Mills LS (2005). Mortality by moonlight: predation risk and the snowshoe hare. *Behavioral Ecology* **16**: 938–944.

Grossmann S, Bauer S, Robinson PN, Vingron M (2007). Improved detection of overrepresentation of Gene-Ontology annotations with parent child analysis. *Bioinformatics* **23**: 3024–3031.

Gui D, Jia K, Xia J, Yang L, Chen J, Wu Y, *et al.* (2013). De novo assembly of the Indo-Pacific humpback dolphin leucocyte transcriptome to identify putative genes involved in the aquatic adaptation and immune response. *PloS ONE* **8**: e72417.

Hansen KD, Brenner SE, Dudoit S (2010). Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Research* **38**: e131–e131.

Hasson T, Walsh J, Cable J, Mooseker MS, Brown SD, Steel KP (1997). Effects of shaker-1 mutations on myosin-VIIa protein and mRNA expression. *Cell Motility and the Cytoskeleton* **37**: 127–138.

Hearing VJ (1999). Biochemical control of melanogenesis and melanosomal organization. *Journal of Investigative Dermatology Symposium Proceedings* **4**: 24–28.

Hegna R (2013). To quiver or to shiver: increased melanization benefits thermoregulation, but reduces warning signal efficacy in the wood tiger moth. *Proceedings of The Royal Society* **280**: 1–9.

Hendry AP, Farrugia TJ, Kinnison MT (2008). Human influences on rates of phenotypic change in wild animal populations. *Molecular Ecology* **17**: 20–29.

Henning F, Jones J, Franchini P, Meyer A (2013). Transcriptomics of morphological color change in polychromatic Midas cichlids. *BMC Genomics* **14**: 1–13.

Hershkovitz V, Sela N, Taha-Salaime L, Liu J, Rafael G, Kessler C, *et al.* (2013). De-novo assembly and characterization of the transcriptome of Metschnikowia fructicola reveals differences in gene expression following interaction with Penicillium digitatum and grapefruit peel. *BMC genomics* **14**: 1–13.

Hodges KE (2000a). The Ecology of Snowshoe Hares in Northern Boreal Forests. In: Ruggiero LF, Aubry KB, Buskirk SW, Koehler GM, Krebs CJ, Mckelvey KS, et al. (eds) *Ecology and conservation of lynx in the United States*, University Press of Colorado, pp 117–162.

Hodges KE (2000b). Ecology of Snowshoe Hares in Southern Boreal and Montane Forests. In: Ruggiero LF, Aubry KB, Buskirk SW, Koehler GM, Krebs CJ, Mckelvey KS, et al. (eds) *Ecology and conservation of lynx in the United States*, University Press of Colorado, pp 163–206.

Hoegh-Guldberg O, Mumby PJ, Hooten AJ, Steneck RS, Greenfield P, Gomez E, *et al.* (2007). Coral reefs under rapid climate change and ocean acidification. *Science* **318**: 1737–1742.

Hoekstra HE (2006). Genetics, development and evolution of adaptive pigmentation in vertebrates. *Heredity* **97**: 222–234.

Hoffmann AA, Willi Y (2008). Detecting genetic responses to environmental change. *Nature Reviews Genetics* **9**: 421–432.

Huang L-H, Kang L (2007). Cloning and interspecific altered expression of heat shock protein genes in two leafminer species in response to thermal stress. *Insect Molecular Biology* **16**: 491–500.

Huszar D, Lynch CA, Fairchild-Huntress V, Dunmore JH, Fang Q, Berkemeier LR, *et al.* (1997). Targeted Disruption of the Melanocortin-4 Receptor Results in Obesity in Mice. *Cell* **88**: 131–141.

Ibáñez A, Polo-Cavia N, López P, Martín J (2014). Honest sexual signaling in turtles: experimental evidence of a trade-off between immune response and coloration in red-eared sliders Trachemys scripta elegans. *Naturwissenschaften*: 1–9.

Inouye DW, Barr B, Armitage KB, Inouye BD (2000). Climate change is affecting altitudinal migrants and hibernating species. *Proceedings of the National Academy of Sciences of the United States of America* **97**: 1630–1633.

IUCN (2008). Lepus americanus. *The IUCN Red List of Threatened Species*. http://maps.iucnredlist.org/map.html?id=41273, Acessed 10 August 2014

Jacobsen E, White C, Emison W (1983). Molting adaptations of Rock Ptarmigan on Amchitka Island, Alaska. *Condor* **85**: 420–426.

Jenni L, Kéry M (2003). Timing of autumn bird migration under climate change: advances in long-distance migrants, delays in short-distance migrants. *Proceedings of The Royal Society of London B* **270**: 1467–1471.

Jin W, Riley RM, Wolfinger RD, White KP, Passador-Gurgel G, Gibson G (2001). The contributions of sex, genotype and age to transcriptional variance in Drosophila melanogaster. *Nature Genetics* **29**: 389–395.

Johnson J (2013). fastq_sync. Available from http://toolshed.g2.bx.psu.edu/repository?repository_id=e572fed50242211a

Kasprzyk A (2011). BioMart: driving a paradigm change in biological data management. *Database : the journal of biological databases and curation* **2011**: 1–3.

Kimura M (1984). *The neutral theory of molecular evolution*. Cambridge University Press.

King M, Wilson A (1975). Evolution at two levels in humans and chimpanzees. *Science* **188**: 107–116.

Kneussel M, Wagner W (2013). Myosin motors at neuronal synapses: drivers of membrane transport and actin dynamics. *Nature reviews Neuroscience* **14**: 233–247.

Kobayashi H, Kromminga A, Dunlop TW, Tychsen B, Conrad F, Suzuki N, *et al.* (2005). A role of melatonin in neuroectodermal-mesodermal interactions: the hair follicle synthesizes melatonin and expresses functional melatonin receptors. *FASEB journal : official publication of the Federation of American Societies for Experimental Biology* **19**: 1710–1712.

Kraemer AC, Adams DC (2014). Predator perception of Batesian mimicry and conspicuousness in a salamander. *Evolution* **68**: 1197–1206.

Langbein L, Rogers MA, Praetzel-Wunder S, Helmke B, Schirmacher P, Schweizer J (2006). K25 (K25irs1), K26 (K25irs2), K27 (K25irs3), and K28 (K25irs4) represent the type I inner root sheath keratins of the human hair follicle. *The Journal of investigative dermatology* **126**: 2377–2386.

Langmead B, Trapnell C, Pop M, Salzberg SL (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology* **10**: R25.1–R25.10.

Larkin JE, Freeman DA, Zucker I (2001). Low ambient temperature accelerates short-day responses in Siberian hamsters by altering responsiveness to melatonin. *Journal of Biological Rhythms* **16**: 76–86.

Larsen PF, Nielsen EE, Williams TD, Hemmer-Hansen J, Chipman JK, Kruhøffer M, *et al.* (2007). Adaptive differences in gene expression in European flounder (Platichthys flesus). *Molecular ecology* **16**: 4674–4683.

Li B, Dewey CN (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**: 1–16.

Lin KK, Chudova D, Hatfield GW, Smyth P, Andersen B (2004). Identification of hair cycle-associated genes from time-course gene expression profile data by using replicate variance. *Proceedings of the National Academy of Sciences of the United States of America* **101**: 15955–15960.

Lin KK, Kumar V, Geyfman M, Chudova D, Ihler AT, Smyth P, *et al.* (2009). Circadian clock genes contribute to the regulation of hair follicle cycling. *PLoS genetics* **5**: e1000573.

Lincoln GA, Clarke IJ, Hut RA, Hazlerigg DG (2006). Characterizing a mammalian circannual pacemaker. *Science* **314**: 1941–1944.

Liu X, Ondek B, Williams DS (1998). Mutant myosin VIIa causes defective melanosome distribution in the RPE of shaker-1 mice. *Nature Genetics* **19**: 117–118.

Logan A, Weatherhead B (1978). Pelage color cycles and hair follicle tyrosinase activity in the Siberian hamster. *The Journal of Investigative Dermatology* **71**: 295–298.

Logan A, Weatherhead B (1980). Post-tyrosinase inhibition of melanogenesis by melatonin in hair follicles in vitro. *The Journal of Investigative Dermatology* **74**: 47–50.

Maechler M, Rousseeuw P, Struyf A, Hubert M, Hornik K (2012). cluster: Cluster Analysis Basics and Extensions. *R package version 1152*.

Marchese C, Felici A, Visco V (2001). Fibroblast growth factor 10 induces proliferation and differentiation of human primary cultured keratinocytes. *Journal of Investigative Dermatology* **116**: 623–628.

Mardis ER (2013). Next-generation sequencing platforms. *Annual Review of Analytical Chemistry* **6**: 287–303.

Margolin G, Khil PP, Kim J, Bellani MA, Camerini-Otero RD (2014). Integrated transcriptome analysis of mouse spermatogenesis. *BMC Genomics* **15**: 1–19.

Marmol V del, Beermann F (1996). Tyrosinase and related proteins in mammalian pigmentation. *FEBS letters* **381**: 165–168.

Maroni G, Wise J, Young J, Otto E (1987). Metallothionein gene duplications and metal tolerance in natural populations of Drosophila melanogaster. *Genetics* **117**: 739–744.

Marra NJ, Romero A, Dewoody JA (2014). Natural selection and the genetic basis of osmoregulation in Heteromyid rodents as revealed by RNA-seq. *Molecular Ecology* **23**: 2699–2711.

Martin M (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet Journal* **17**: 10–12.

Martin JA, Wang Z (2011). Next-generation transcriptome assembly. *Nature Reviews Genetics* **12**: 671–682.

Matthee C, Van Vuuren B, Bell D, Robinson T (2004). A Molecular Supermatrix of the Rabbits and Hares (Leporidae) Allows for the Identification of Five Intercontinental Exchanges During the Miocene. *Systematic Biology* **53**: 433–447.

Mazar J, DeYoung K, Khaitan D, Meister E, Almodovar A, Goydos J, *et al.* (2010). The regulation of miRNA-211 expression and its role in melanoma cell invasiveness. *PloS ONE* **5**: e13779.

McCarthy DJ, Chen Y, Smyth GK (2012). Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Research* **40**: 4288–4297.

McDade SS, Henry AE, Pivato GP, Kozarewa I, Mitsopoulos C, Fenwick K, *et al.* (2012). Genome-wide analysis of p63 binding sites identifies AP-2 factors as co-regulators of epidermal differentiation. *Nucleic Acids Research* **40**: 7190–7206.

Melo-Ferreira J, Boursot P, Carneiro M, Esteves PJ, Farelo L, Alves PC (2012). Recurrent introgression of mitochondrial DNA among hares (Lepus spp.) revealed by species-tree inference and coalescent simulations. *Systematic Biologyology* **61**: 367–381.

Melo-Ferreira J, Boursot P, Randi E, Kryukov A, Suchentrunk F, Ferrand N, *et al.* (2007). The rise and fall of the mountain hare (Lepus timidus) during Pleistocene glaciations: expansion and retreat with hybridization in the Iberian Peninsula. *Molecular Ecology* **16**: 605–618.

Melo-Ferreira J, Seixas FA, Cheng E, Mills LS, Alves PC (2014). The hidden history of the snowshoe hare, Lepus americanus: extensive mitochondrial DNA introgression inferred from multilocus genetic variation. *Molecular Ecology*: 4617–4630.

Meredith AL, Wiler SW, Miller BH, Takahashi JS, Fodor AA, Ruby NF, *et al.* (2006). BK calcium-activated potassium channels regulate circadian behavioral rhythms and pacemaker output. *Nature Neuroscience* **9**: 1041–1049.

Merkin J, Russell C, Chen P, Burge C (2012). Evolutionary dynamics of gene and isoform regulation in Mammalian tissues. *Science* **338**: 1593–1599.

Metzker ML (2010). Sequencing technologies - the next generation. *Nature Reviews Genetics* **11**: 31–46.

Miller JR, Koren S, Sutton G (2010). Assembly algorithms for next-generation sequencing data. *Genomics* **95**: 315–327.

Mills LS, Zimova M, Oyler J, Running S, Abatzoglou JT, Lukacs PM (2013). Camouflage mismatch in seasonal coat color due to decreased snow duration. *Proceedings of the National Academy of Sciences of the United States of America* **110**: 7360–7365.

Mortazavi A, Williams BA, Mccue K, Schaeffer L, Wold B (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods* **5**: 621–628.

Nachman MW, Hoekstra HE, D'Agostino SL (2003). The genetic basis of adaptive melanism in pocket mice. *Proceedings of the National Academy of Sciences of the United States of America* **100**: 5268–5273.

Nagalakshmi U, Wang Z, Waern K, Shou C (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320**: 1344–1349.

Nakamura M, Tobin DJ, Richards-Smith B, Sundberg JP, Paus R (2002). Mutant laboratory mice with abnormalities in pigmentation: annotated tables. *Journal of dermatological science* **28**: 1–33.

Nunome M, Kinoshita G, Tomozawa M, Torii H, Matsuki R, Yamada F, *et al.* (2014). Lack of association between winter coat colour and genetic population structure in the Japanese hare, Lepus brachyurus (Lagomorpha: Leporidae). *Biological Journal of the Linnean Society* **111**: 761–776.

Orr HA (2005). The genetic theory of adaptation: a brief history. *Nature Reviews Genetics* **6**: 119–127.

Ortolani A (1999). Spots, stripes, tail tips and dark eyes: predicting the function of carnivore colour patterns using the comparative method. *Biological Journal of the Linnean Society* **67**: 433–476.

Papakostas S, Vøllestad LA, Bruneaux M, Aykanat T, Vanoverbeke J, Ning M, *et al.* (2014). Gene pleiotropy constrains gene expression changes in fish adapted to different thermal conditions. *Nature Communications* **5**: 1–9.

Parchem RJ, Perry MW, Patel NH (2007). Patterns on the insect wing. *Current Opinion in Genetics & Development* **17**: 300–308.

Parmesan C (2006). Ecological and Evolutionary Responses to Recent Climate Change. *Annual Review of Ecology, Evolution, and Systematics* **37**: 637–669.

Parmesan C, Yohe G (2003). A globally coherent fingerprint of climate change impacts across natural systems. *Nature* **421**: 37–42.

Paus R, Cotsarelis G (1999). The biology of hair follicles. *The New England Journal of Medicine* **341**: 491–497.

Pederson G, Gray S, Woodhouse C (2011). The unusual nature of recent snowpack declines in the North American Cordillera. *Science* **333**: 332–335.

Preitner N, Damiola F, Zakany J, Duboule D, Albrecht U, Schibler U (2002). The Orphan Nuclear Receptor REV-ERBα Controls Circadian Transcription within the Positive Limb of the Mammalian Circadian Oscillator. *Cell* **110**: 251–260.

Pritchard C, Hsu L, Delrow J, Nelson PS (2001). Project normal: defining normal variance in mouse gene expression. *Proceedings of the National Academy of Sciences of the United States of America* **98**: 13266–13271.

Protas ME, Patel NH (2008). Evolution of coloration patterns. *Annual Review of Cell and Developmental Biology* **24**: 425–446.

Quinn G, Keough M (2002). *Experimental design and data analysis for biologists*, 1st edn. Cambridge University Press: New York.

Raveh E, Cohen S, Levanon D, Groner Y, Gat U (2005). Runx3 is involved in hair shape determination. *Developmental dynamics : an official publication of the American Association of Anatomists* **233**: 1478–1487.

Raveh E, Cohen S, Levanon D, Negreanu V, Groner Y, Gat U (2006). Dynamic expression of Runx1 in skin affects hair structure. *Mechanisms of development* **123**: 842–50.

Réale D, McAdam AG, Boutin S, Berteaux D (2003). Genetic and plastic responses of a northern mammal to climate change. *Proceedings of The Royal Society of London B* **270**: 591–596.

Reppert SM, Weaver DR (2002). Coordination of circadian timing in mammals. *Nature* **418**: 935–941.

Reynolds AP, Richards G, Iglesia B, Rayward-Smith VJ (2006). Clustering Rules: A Comparison of Partitioning and Hierarchical Clustering Algorithms. *Journal of Mathematical Modelling and Algorithms* **5**: 475–504.

Robinson MD, McCarthy DJ, Smyth GK (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**: 139–140.

Robinson MD, Oshlack A (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology* **11**: 1–9.

Robinson MD, Smyth GK (2007). Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics* **23**: 2881–2887.

Rowland W, Baube C, Horan T (1991). Signalling of sexual receptivity by pigmentation pattern in female sticklebacks. *Animal Behaviour* **42**: 243–249.

Rust C (1965). Hormonal control of pelage cycles in the short tailed weasel (Mustela erminea bangsi). *General and Comparative Endocrinology* **5**: 222–231.

Schlake T (2005a). FGF signals specifically regulate the structure of hair shaft medulla via IGF-binding protein 5. *Development* **132**: 2981–2990.

Schlake T (2005b). Segmental Igfbp5 expression is specifically associated with the bent structure of zigzag hairs. *Mechanisms of Development* **122**: 988–997.

Schlake T, Beibel M, Weger N, Boehm T (2004). Major shifts in genomic activity accompany progression through different stages of the hair cycle. *Gene Expression Patterns : GEP* **4**: 141–152.

Seixas F (2012). The Speciation History of North American Hares (Lepus spp .): Divergence with gene flow? Universidade do Porto.

Severaid JH (1945). Pelage Changes in the Snowshoe Hare. *Journal of Mammalogy* **26**: 41–63.

Shendure J, Ji H (2008). Next-generation DNA sequencing. *Nature Biotechnology* **26**: 1135–1145.

Shiraki T, Kondo S, Katayama S, Waki K, Kasukawa T, Kawaji H, *et al.* (2003). Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proceedings of the National Academy of Sciences of the United States of America* **100**: 15776–15781.

Singhal S (2013). De novo transcriptomic analyses for non-model organisms: an evaluation of methods across a multi-species data set. *Molecular Ecology Resources* **13**: 403–416.

Skoglund P, Höglund J (2010). Sequence polymorphism in candidate genes for differences in winter plumage between Scottish and Scandinavian Willow Grouse (Lagopus lagopus). *PloS ONE* **5**: e10334.

Slominski A, Chassalevris N, Mazurkiewicz J, Maurer M, Paus R (1994). Murine skin as a target for melatonin bioregulation. *Experimental Dermatology* **3**: 45–50.

Slominski A, Fischer TW, Zmijewski M a, Wortsman J, Semak I, Zbytek B, *et al.* (2005). On the role of melatonin in skin physiology and pathology. *Endocrine* **27**: 137–148.

Slominski A, Paus R (1993). Melanogenesis is coupled to murine anagen: toward new concepts for the role of melanocytes and the regulation of melanogenesis in hair growth. *Journal of Investigative Dermatology* **101**: 90S–97S.

Slominski R, Reiter RJ, Schlabritz-Loutsevitch N, Ostrom RS, Slominski AT (2012). Melatonin membrane receptors in peripheral tissues: distribution and functions. *Molecular and Cellular Endocrinology* **351**: 152–166.

Slominski A, Tobin DJ, Shibahara S, Wortsman J (2004). Melanin pigmentation in mammalian skin and its hormonal regulation. *Physiological Reviews* **84**: 1155–1228.

Solomon S, Plattner G-K, Knutti R, Friedlingstein P (2009). Irreversible climate change due to carbon dioxide emissions. *Proceedings of the National Academy of Sciences of the United States of America* **106**: 1704–1709.

Spencer JD, Schallreuter KU (2009). Regulation of pigmentation in human epidermal melanocytes by functional high-affinity beta-melanocyte-stimulating hormone/melanocortin-4 receptor signaling. *Endocrinology* **150**: 1250–1258.

Stapley J, Reger J, Feulner PGD, Smadja C, Galindo J, Ekblom R, *et al.* (2010). Adaptation genomics: the next generation. *Trends in Ecology & Evolution* **25**: 705–712.

Steijger T, Abril JF, Engström PG, Kokocinski F, Akerman M, Alioto T, *et al.* (2013). Assessment of transcript reconstruction methods for RNA-seq. *Nature Methods* **10**: 1177–1184.

Steiner CC, Weber JN, Hoekstra HE (2007). Adaptive variation in beach mice produced by two interacting pigmentation genes. *PLoS Biology* **5**: e219.

Stenn KS, Paus R (2001). Controls of hair follicle cycling. *Physiological Reviews* **81**: 449–494.

Stoner C (2003). The adaptive significance of coloration in lagomorphs. *Biological Journal of the Linnean Society* **79**: 309–328.

Stoughton RB (2005). Applications of DNA microarrays in biology. *Annual Review of Biochemistry* **74**: 53–82.

Stuart-Fox D, Whiting M, Moussalli A (2006). Camouflage and colour change: antipredator responses to bird and snake predators across multiple populations in a dwarf chameleon. *Biological Journal of the Linnean Society* **88**: 437–446.

Swope VB, Abdel-Malek Z, Kassem LM, Nordlund JJ (1991). Interleukins 1alpha and 6 and Tumor Necrosis Factor-alpha Are Paracrine Inhibitors of Human Melanocyte Proliferation and Melanogenesis. *Journal of Investigative Dermatology* **96**: 180–185.

Team RC (2014). R: A language and environment for statistical computing.

Thomas CD, Cameron A, Green RE, Bakkenes M, Beaumont LJ, Collingham YC, *et al.* (2004). Extinction risk from climate change. *Nature* **427**: 145–148.

Tibshirani R, Walther G, Hastie T (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B* **63**: 411–423.

Uzawa K, Grzesik W, Nishiura T, Kuznetsov SA, Robey PG, Brenner DA, *et al.* (1999). Differential Expression of Human Lysyl Hydroxylase Genes, Lysine Hydroxylation, and Cross-Linking of Type I Collagen During Osteoblastic Differentiation In. *Journal of Bone and Mineral Research* **14**: 1272–1280.

Våge DI, Fuglei E, Snipstad K, Beheim J, Landsem VM, Klungland H (2005). Two cysteine substitutions in the MC1R generate the blue variant of the Arctic fox (Alopex lagopus) and prevent expression of the white winter coat. *Peptides* **26**: 1814–1817.

Velculescu V, Zhang L, Vogelstein B, Kinzler K (1995). Serial analysis of gene expression. *Science* **270**: 484–487.

Vellichirammal NN, Zera AJ, Schilder RJ, Wehrkamp C, Riethoven J-JM, Brisson J a (2014). De novo transcriptome assembly from fat body and flight muscles transcripts to identify morph-specific gene expression profiles in Gryllus firmus. *PloS ONE* **9**: e82129.

Vijay N, Poelstra JW, Künstner A, Wolf JBW (2013). Challenges and strategies in transcriptome assembly and differential gene expression quantification. A comprehensive in silico assessment of RNA-seq experiments. *Molecular Ecology* **22**: 620–634.

Voisey J, Daal A Van (2002). Agouti: from mouse to man, from skin to fat. *Pigment Cell Research* **15**: 10–18.

Vrieling H, Duhl DM, Millar SE, Miller K a, Barsh GS (1994). Differences in dorsal and ventral pigmentation result from regional expression of the mouse agouti gene. *Proceedings of the National Academy of Sciences of the United States of America* **91**: 5667–5671.

Wang Z, Gerstein M, Snyder M (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics* **10**: 57–63.

Wang X-W, Luan J-B, Li J-M, Bao Y-Y, Zhang C-X, Liu S-S (2010). De novo characterization of a whitefly transcriptome and analysis of its gene expression during development. *BMC Genomics* **11**: 1–10.

Warnes GR, Bolker B, Bonebakker L, Gentleman R, Liaw W, Lumley T, *et al.* (2014). ggplots: Various R programming tools for plotting data.

Whitehead A, Crawford DL (2005). Variation in tissue-specific gene expression among natural populations. *Genome Biology* **6**: R13.1–R13.14.

Whitney AR, Diehn M, Popper SJ, Alizadeh A a, Boldrick JC, Relman D a, *et al.* (2003). Individuality and variation in gene expression patterns in human blood. *Proceedings of the National Academy of Sciences of the United States of America* **100**: 1896–1901.

Wolf JBW (2013). Principles of transcriptome analysis and gene expression quantification: an RNA-seq tutorial. *Molecular Ecology Resources* **13**: 559–572.

Wu X, Bowers B, Rao K, Wei Q, Hammer JA (1998). Visualization of melanosome dynamics within wild-type and dilute melanocytes suggests a paradigm for myosin V function In vivo. *The Journal of Cell Biology* **143**: 1899–1918.

Yang Y, Smith SA (2013). Optimizing de novo assembly of short-read RNA-seq data for phylogenomics. *BMC genomics* **14**: 1–11.

Yeo G, Farooqi I, Aminian S (1998). A frameshift mutation in MC4R associated with dominantly inherited human obesity. *Nature Genetics* **20**: 111–112.

Yin L, Wu N, Lazar M a (2010). Nuclear receptor Rev-erbalpha: a heme receptor that coordinates circadian rhythm and metabolism. *Nuclear Receptor Signaling* **8**: 1–6.

Zhou Y, Chen L, Fan X, Bian Y (2014). De novo assembly of Auricularia polytricha transcriptome using Illumina sequencing for gene discovery and SSR marker identification. *PloS ONE* **9**: e91740.

Zimova M, Mills L (2014). Snowshoe hares display limited phenotypic plasticity to mismatch in seasonal camouflage. *Proceedings of The Royal Society B* **281**: 1–19.

Appendix

The document *APPENDIX_1.xlsx* in Excel Microsoft Office 2010 .xlsx format provided is referred in the text as Appendix 1 and contains the following list of tables:

Table S1 - List of differentially expressed Trinity components and respective ENSEMBL gene code and associated gene name resultant from annotation with blastx against *Oryctolagus cuniculus* protein database. Results here are shown at the gene level.

Table S2 - List of trinity components included in each cluster resultant from partinioning clustering with k=3.

Table S3 - Results for Gene Ontology (GO) enrichment analysis with the genes differentially expressed between "brown" and "white" stage. Only significantly enrichment terms (FDR < 0.05) are shown. C - Cellular Component; M - Molecular Function; B - Biological Process.

Table S4 - Results for Gene Ontology (GO) enrichment analysis with the genes differentially expressed between "brown" and "molting" stage. Only significantly enrichment terms (FDR < 0.05) are shown. C - Cellular Component; M - Molecular Function; B - Biological Process.

Table S5 - Results for Gene Ontology (GO) enrichment analysis with the genes differentially expressed between "white" and "molting" stage. Only significantly enrichment terms (FDR < 0.05) are shown. C - Cellular Component; M - Molecular Function; B - Biological Process.